Predictive Modeling and Big Data Technology on Kickstarter Success

Introduction:

Compared to traditional financing, more startups are using Internet platforms to appeal to crowd investors, which is Crowdfunding (CF). While external funding has become not prevalent, many entrepreneurs have used CF to reach various financial sources. Kickstarter1 is one of the largest crowdfunding platforms and includes projects from multiple categories. The dataset is structured and is accessible from 2009 to 2019; it includes a main table of projects, with side tables including reward, location, category, comments, creators, reward items, and Livestream. With over 40M observations, the overall size of the dataset is over 10GB. Through analyzing the data, we decide to pull some model-free insights using SQL (Figure 1), and build a machine learning model to predict whether a project will succeed. We aim to give entrepreneurial companies general ideas of what kind of project is more profitable and feasible, increasing their success rate.

Data Wrangling:

We built an ER model (Figure 10) to visualize all tables' relationships. After carefully analyzing the ER model and all variables across all tables, we decided to join the category table to our main table, since category can be a possible important feature, and use the joined table for future analysis.

In our data exploring phase, we first take a look at the distribution of our target variable (column 'state'), which has six different categories to describe the funding status of each project, the six categories are 'successful', 'canceled', 'failed', 'suspended', 'purged' and 'live.' To align with our goal of predicting whether a project can be successful, we encode 'successful' as 1 and the rest 5 states as 0. Then from the bar graph (Figure 2), we can quickly tell our target variable is not very balanced, so we oversample the data to balance out our target variable.

Next, we perform statistical analysis (Figure 3) to our numerical variables to get a basic idea of what these variables look like. We also drew density plots to visualize the data distribution (Figure 4) of these numerical variables and found that most are incredibly right-skewed and need transformation. This transformation is quite challenging since we decided to use log transformation after testing several data transformation techniques. Still, we have a significant number of 0s, which will be minus infinity after the log. So we add 0.01 to all values to avoid taking log zero while minimizing the effect of adding an arbitrary value. Besides, although log transformation did reduce our data skewness to some extent (Figure 5), there is still room for improvement, so advanced techniques may still be helpful.

Then, we noticed our date-related data were in Unix format. We had to convert those back to the date and generate a new feature called project_duration by calculating the difference between the project end date from the start date since project duration can be a relevant feature. Also, we created dummy variables for our categorical variables to ensure they can be included in the model.

Insights & Analytics:

After performing the EDA in python 3, we found most of the projects are carried out in the United States, which counts more than 76%, followed by the United Kingdom and Canada (Figure 6). For the industry the company invests in, most of the investments carry on the top 10 sectors; film & video is the most popular investment industry. We also notice some industries have a higher chance of success, like music, comics, and theater. Meanwhile, some industries, like technology, fashion and food, tend to have a higher chance of failure (Figure 7). On top of this, we also build a heatmap to check the correlations among all the variables we have. With the

heatmap, we found a highly positive relationship between the pledged amount and the number of backers, which means as the pledged amount goes higher, the number of backers will also increase. Besides this, most of the columns have weak connections between each other. In addition, we created an interactive dashboard for further analysis in the US market (Reference 3). After the initial EDA and data cleaning, we import our dataset into S3 and use PySpark to build logistic regression and random forest machine learning model. Before creating the model, we split our data into two parts, 70% of the data for training and 30% for testing. We also use hyperparameter tuning to determine the best parameter in each machine-learning model. After this, we perform two different evaluation ways to evaluate our model: AUC (Area under the ROC Curve) and accuracy. From the model we train, we realize an essential attribute is whether the project receives the spotlight (Figure 8). On the other hand, which country the project shows the least important feature of the model (Figure 9). With this machine learning model, we can use it to predict whether or not the project will succeed in the future.

After we built two different models and performed the evaluations in PySpark, we noticed the performance of random forest was slightly better than logistic regression. Therefore, we decided to choose a random forest as our machine-learning model. Besides, the random forest machine learning model has the following advantages: (1) It can handle large datasets efficiently; (2) It reduces overfitting in decision trees and helps to improve accuracy.

Discussion:

Based on our analysis above, we provided insights to entrepreneurial companies on what kind of project is more profitable and feasible. For instance, companies who plan to initiate a tech

project may need to lower their expectation of success and aim for a smaller goal amount. In contrast, companies that plan to begin a music project can propose a higher goal amount.

Besides, we can also provide insights to backers. Based on the information we get from the machine learning model, whether or not the project gets the spotlight is critical to the success rate. Therefore, we recommend paying attention to projects with higher goals(pledged amount), as which countries the project carries on seems not directly relevant to the success rate. Hence, we recommend that investors focus on the project in the top 3 countries, the United States, United Kingdom, and Canada since these three count for more than 90% of the project.

Appendix-Exhibits

1. top 5 country that have the most successful crowdfunding project

project_count	country
123035	US
14792	GB
5240	CA
2506	AU
1373	DE

2. top 10 category that have the most successful crowdfunding project

project_count	category
28120	music
25622	film & video
15806	games
14065	publishing
13431	art
12767	design
7484	technology
7325	comics
7062	fashion
6940	theater

3. avg amount pledged by country (top 5)

avg_amount	country
20436.82	CH
19192.1	AT
17853.61	HK
15183.04	JP
12891.37	FR

4.avg amount pledged by category

Result Grid | Filter Rows:

avg_amount	category
27546.29	design
24149.83	games
22501.5	technology
6913.88	comics
6319.98	film & video

Figure 1:SQL

EDA

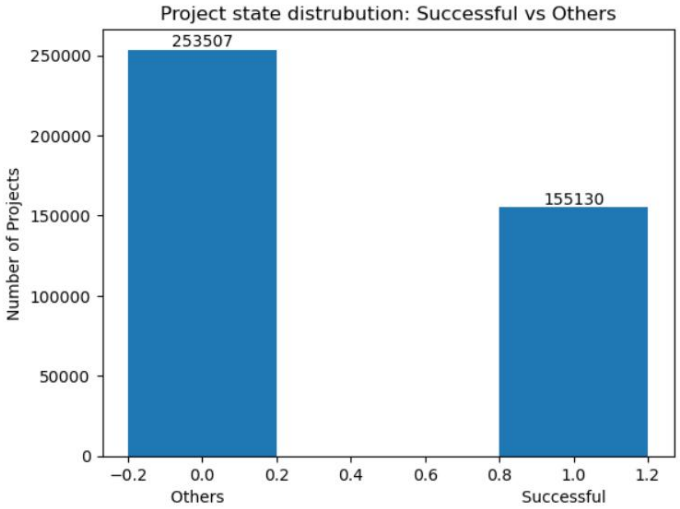

Figure 2: Target Variable Distribution

	goal	usd_pledged	backers_count	comments_count	updates_count
count	408636.00	408636.00	408636.00	408636.00	408636.00
mean	50277.37	9932.61	115.33	40.32	4.91
std	1170800.36	92920.22	913.46	1115.13	9.53
min	0.01	0.00	0.00	0.00	0.00
25%	2000.00	44.76	2.00	0.00	0.00
50%	5200.00	746.00	14.00	0.00	1.00
75%	16000.00	4468.34	62.00	3.00	6.00
max	100000000.00	20338986.27	219382.00	393425.00	412.00

Figure 3: numerical variable statistical analysis

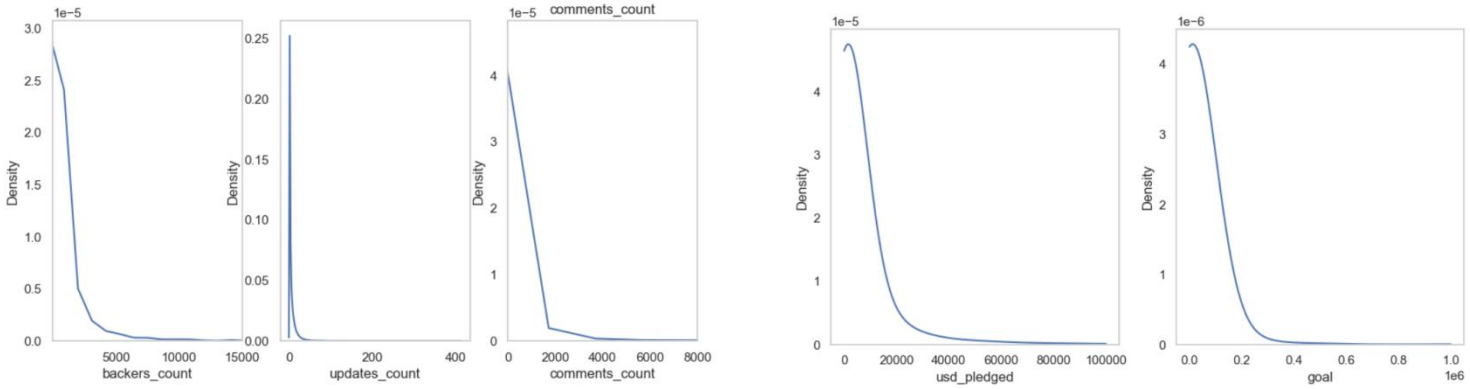

Figure 4:numerical variable density plot

Skewness before log transformation：

```
goal              70.48
usd_pledged       74.55
backers_count     80.31
comments_count   161.22
updates_count      5.39
dtype: float64
```

Skewness after log transformation：

```
goal             -0.11
usd_pledged      -4.52
backers_count    -0.91
comments_count    0.69
updates_count    -0.13
dtype: float64
```

Figure 5:numerical variable skewness before and after transformation

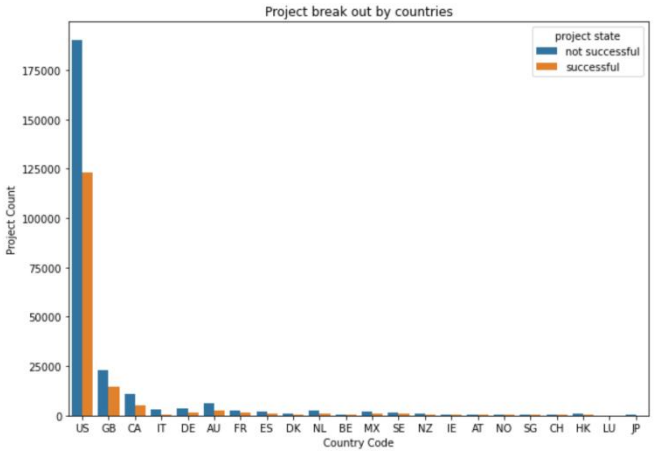

Figure 6

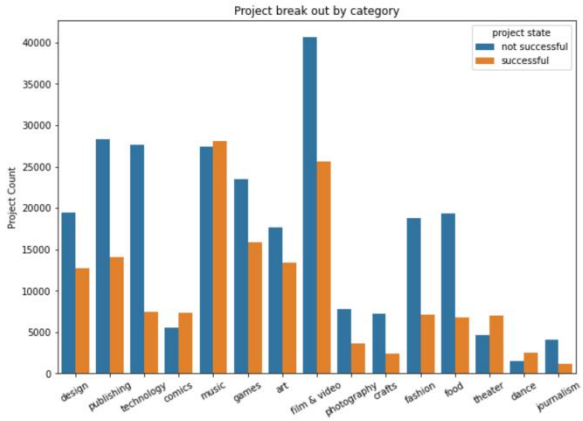

Figure 7

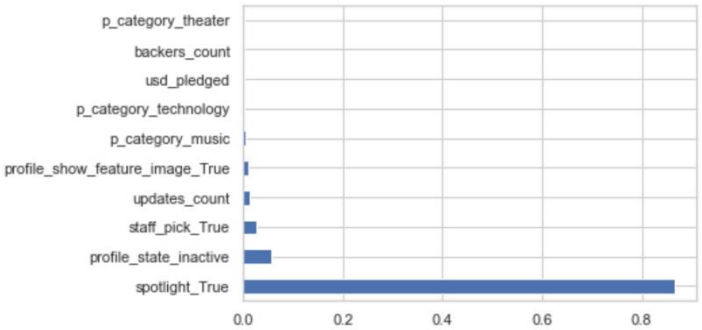

Figure 8: most important features

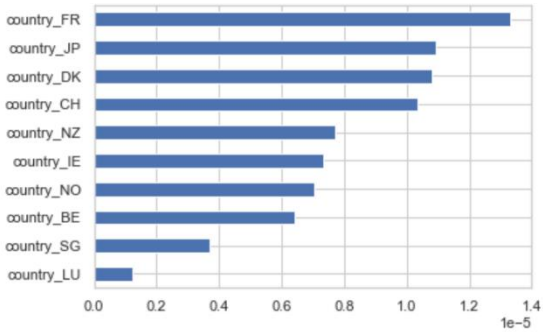

Figure 9: least important features

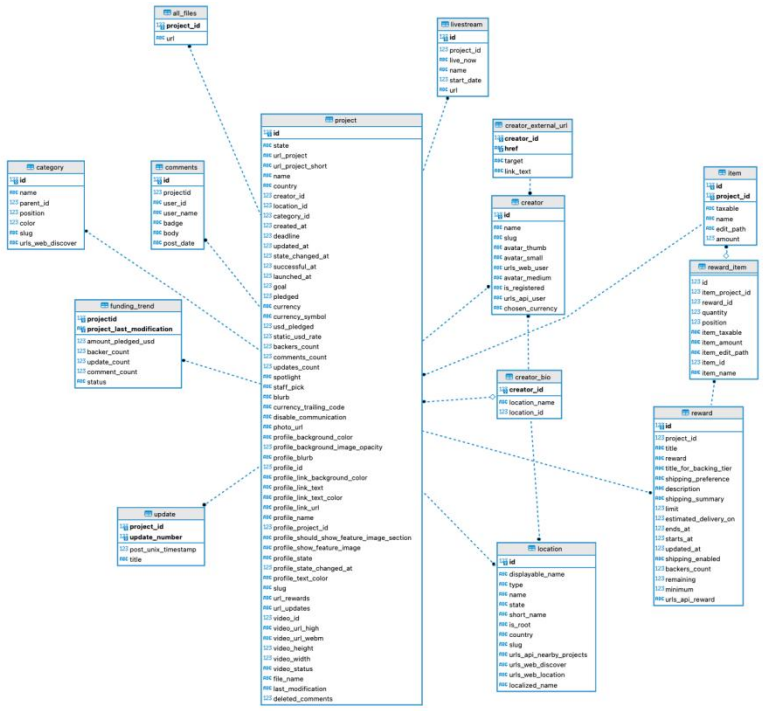

Figure 10: ER model

Reference website and data:

1. Dataset:https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/EOYBXM
2. LI, GUAN-CHENG, 2019, "Kickstarter Structured Relational Database", https://doi.org/10.7910/DVN/EOYBXM
3. Tableau for further exploration in the US market: https://public.tableau.com/app/profile/rebecca.li1104/viz/kickstarter_workbook/kickstarter_story?publish=yes