BIA 652 Project

Preliminary Result

Crowdfunding Successful Forecast

Group 10: Chenkai Hang

Haihan Hu

Haodong Zhao

Project introduction

Kickstarter is the world's largest funding platform for creative projects. We scraped over 2300 instances from Kuckstarter.com by using web mining. We want to build a classification model to predict if a project would success.

Data cleanup

Following screenshot is part of our scraped dataset

	ID	goal	pledged	currency	deadline	created_at	launched_at	backers_coun	percent_fund	loc_country	cate_name	success_or_fail	videoCount	imageCount
0	1193151603	780	262	USD	1533677994	1491347901	1531949994	18	33.5897436	US	Art	0	0	5
1	1428839533	12000	12593	USD	1542178740	1537747420	1539788433	209	104.941667	US	Childrenswea	1	1	15
2	761677304	10000	3013	USD	1542179338	1522598214	1539583738	38	30.13	US	Children's Bo	0	0	2
3	289600410	9112	9184.32	AUD	1542179617	1537656570	1539584017	57	100.793679	AU	Children's Bo	1	1	7
4	1787520409	625	717	USD	1542182340	1538711556	1540253605	23	114.72	US	Comedy	1	1	3
5	1949755547	15000	25915	USD	1542182400	1540175448	1540267229	459	172.766667	US	Graphic Nove	1	1	27
6	957556678	5000	5211	GBP	1542183279	1532095457	1539587679	43	104.22	GB	Art	1	0	16
7	1395473762	10000	12	USD	1542183830	1536942119	1536996230	3	0.12	US	Metal	0	0	0
8	660581565	1600000	3560	HKD	1542184177	1536580245	1538292577	4	0.2225	HK	DIY Electroni	0	0	14
9	975345120	20000	4998	USD	1542185242	1538509385	1539589642	41	24.99	US	Thrillers	0	0	16
10	693384120	1000	1	EUR	1542185245	1529252864	1539071245	1	0.1	IT	Music	0	0	0
11	1304569423	4500	9308	USD	1542193738	1536006702	1539598138	222	206.844444	US	Graphic Desi	1	1	25
12	1081358036	30000	703.32	AUD	1542194488	1539662530	1540030888	5	2.3444	AU	Documentary	0	0	0
13	2078937987	2000	509	EUR	1542195119	1536263153	1539599519	15	25.45	FR	Typography	0	0	10
14	344948511	35851	36568.71	SEK	1542196027	1536683538	1539600427	126	102.001925	SE	Fiction	1	1	10

In this dataset, we have some different currency, so we cannot directly use the values in 'goal' and 'pledged' columns to build a model. We wrote a currency conversion function to change all values in 'goal' and 'pledged' columns to USD. Besides, the time in columns 'deadline', 'created_at' and 'launched_at' are Unix timestamp format, we transformed them to yyyy-mm-dd format. Then after drop NA value, we got the following new dataset.

	ID	goal	pledged	currency	deadline	created_at	launched_at	backers_count	percent_funded	loc_country	cate_name	success_or_fail	videoCount	imageCount
0	1.19E+09	780	262	USD	2018-08-0	2017-04-04	2018-07-18	18	33.58974359	US	Art	0	0	5
1	1.43E+09	12000	12593	USD	2018-11-1	2018-09-23	2018-10-17	209	104.9416667	US	Childrenswear	1	1	15
2	7.62E+08	10000	3013	USD	2018-11-1	2018-04-01	2018-10-15	38	30.13	US	Children's Books	0	0	2
3	2.9E+08	6470	6521	USD	2018-11-1	2018-09-22	2018-10-15	57	100.7936787	AU	Children's Books	1	1	7
4	1.79E+09	625	717	USD	2018-11-1	2018-10-04	2018-10-22	23	114.72	US	Comedy	1	1	3
5	1.95E+09	15000	25915	USD	2018-11-1	2018-10-21	2018-10-23	459	172.7666667	US	Graphic Novels	1	1	27
6	9.58E+08	6500	6774	USD	2018-11-1	2018-07-20	2018-10-15	43	104.22	GB	Art	1	0	16
7	1.4E+09	10000	12	USD	2018-11-1	2018-09-14	2018-09-15	3	0.12	US	Metal	0	0	0
8	6.61E+08	208000	463	USD	2018-11-1	2018-09-10	2018-09-30	4	0.2225	HK	DIY Electronics	0	0	14
9	9.75E+08	20000	4998	USD	2018-11-1	2018-10-02	2018-10-15	41	24.99	US	Thrillers	0	0	16
10	6.93E+08	1210	1	USD	2018-11-1	2018-06-17	2018-10-09	1	0.1	IT	Music	0	0	0
11	1.3E+09	4500	9308	USD	2018-11-1	2018-09-03	2018-10-15	222	206.8444444	US	Graphic Design	1	1	25

Data Analysis

For preliminary analysis, we set column ‘success_or_fail’ as y, set columns ‘videoCount’ and ‘imageCount’ as x. Then use 75% of data as training data and 25% of data as validation data. Then we train Linear Discriminant Analysis model and Logistic Regression model by using training data and then predict by using validation data. Following are the score of LDA model and Logistic Regression model:

```
/usr/local/bin/python3.7 /Users/haodong/Desktop/BIA652/project.py

LDA model
0.7606112054329371

Logistic regression
0.8098471986417657

Process finished with exit code 0
```

Next Step

The dataset still has some non-number attributes, we have to continue cleanup the data. For example, set dummy value for those non-number attributes.

Currently, we only use three variables in our models, for future analysis, we will use more attributes as variables in our models.

After we have enough variables, we will build different models to test the predicted accuracy and then select the best model or ensemble the models.