KICKSTARTER PROJECTS

STEVEN NGUYEN
RYAN GIRON
JINSEOK LEE
ANGELICA SIMITYAN

Question

Dataset

Kickstarter Projects

More than 300,000 kickstarter projects

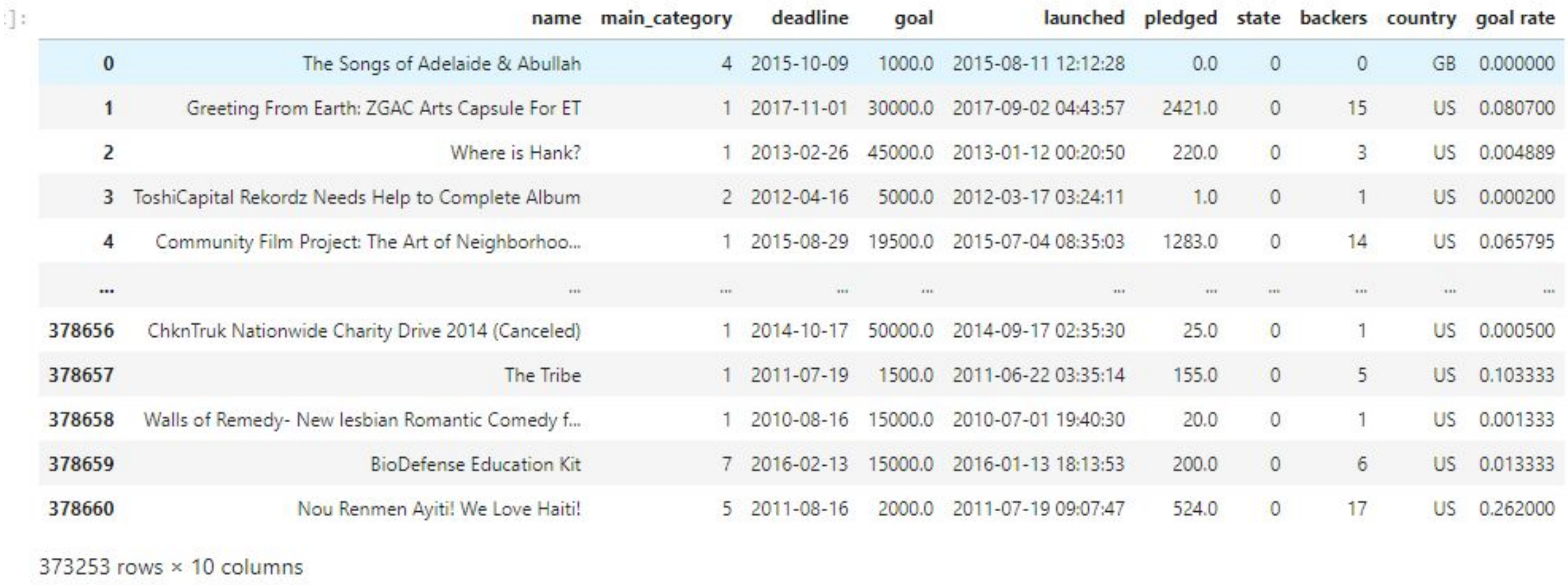
Mickaël Mouillé • updated 4 years ago (Version 7)

Our Dataset: https://www.kaggle.com/kemical/kickstarter-projects

Our Topic: For our project, we are planning to predict the success rate of kickstart projects based on the data given to us in the dataset. Some techniques we are planning to use to achieve this are by using k-nearest neighbor, logistic regression, and comparing both of these results.

Data Cleaning

- Added new column called, “Goal rate”, which shows a rate of how much money a project got towards their set goal amount
  - **Reason:** to see if there is a correlation between goal rate and the success rate
- Removed columns that are not meaningful in our EDA or model
- Regrouped “Country” column into 5 countries with the most projects
  - **Reason:** to avoid possible errors occurring since the data set has many countries that had less than 1% of projects
- Converted non-numeric columns into numeric columns
  - **Reason:** for our model to be able to train itself with numerical data set
- Changed our classification column called “state” to be either 0 or 1
  - 0 if it “failed” or was “canceled”
  - 1 if it was “successful” or “live”
  - **Reason:** for our model to be able to train itself with numerical data set

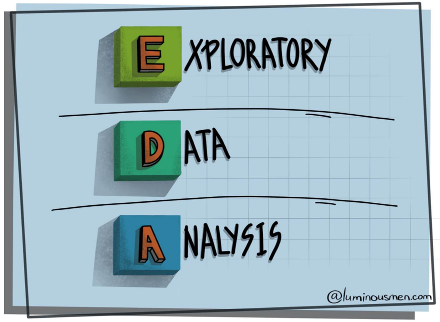


]:

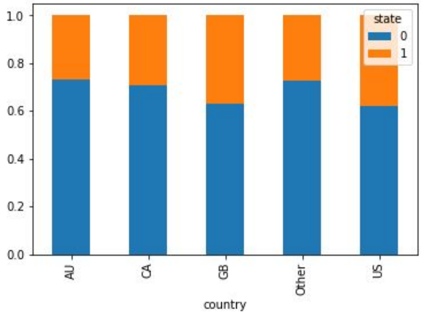
	name	main_category	deadline	goal	launched	pledged	state	backers	country	goal rate
0	The Songs of Adelaide & Abullah	4	2015-10-09	1000.0	2015-08-11 12:12:28	0.0	0	0	GB	0.000000
1	Greeting From Earth: ZGAC Arts Capsule For ET	1	2017-11-01	30000.0	2017-09-02 04:43:57	2421.0	0	15	US	0.080700
2	Where is Hank?	1	2013-02-26	45000.0	2013-01-12 00:20:50	220.0	0	3	US	0.004889
3	ToshiCapital Rekordz Needs Help to Complete Album	2	2012-04-16	5000.0	2012-03-17 03:24:11	1.0	0	1	US	0.000200
4	Community Film Project: The Art of Neighborhoo...	1	2015-08-29	19500.0	2015-07-04 08:35:03	1283.0	0	14	US	0.065795
...	...	...	...	...	...	...	...	...	...	...
378656	ChknTruk Nationwide Charity Drive 2014 (Canceled)	1	2014-10-17	50000.0	2014-09-17 02:35:30	25.0	0	1	US	0.000500
378657	The Tribe	1	2011-07-19	1500.0	2011-06-22 03:35:14	155.0	0	5	US	0.103333
378658	Walls of Remedy- New lesbian Romantic Comedy f...	1	2010-08-16	15000.0	2010-07-01 19:40:30	20.0	0	1	US	0.001333
378659	BioDefense Education Kit	7	2016-02-13	15000.0	2016-01-13 18:13:53	200.0	0	6	US	0.013333
378660	Nou Renmen Ayiti! We Love Haiti!	5	2011-08-16	2000.0	2011-07-19 09:07:47	524.0	0	17	US	0.262000

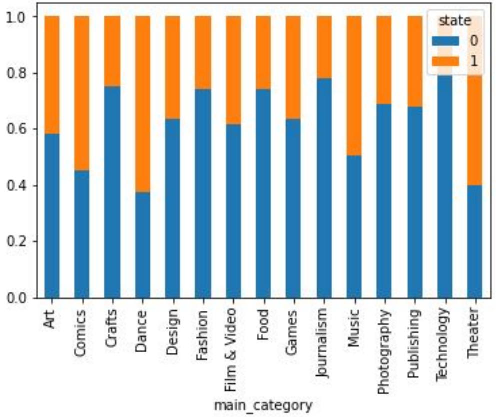
373253 rows × 10 columns

Exploratory Data Analysis (EDA)



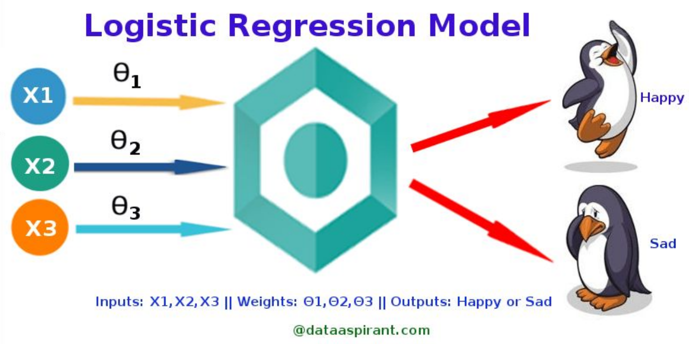
- Made two bar graphs:
  1. Showing the proportion of failed projects based on “main_category”
  2. Showing the proportion of successful projects based on “main_category”
- Created a cross tab of “country” / “main_category” with “state” and made it into a stacked bar graph to show the success rate
- Discovered:
  - Most countries had a similar success to fail ratio
  - Technology had the highest fail rate in “main_category”
  - Theater/Dance had the highest success rate in “main_category”





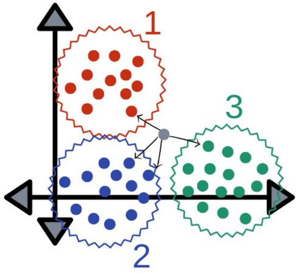
Logistic Regression Analysis

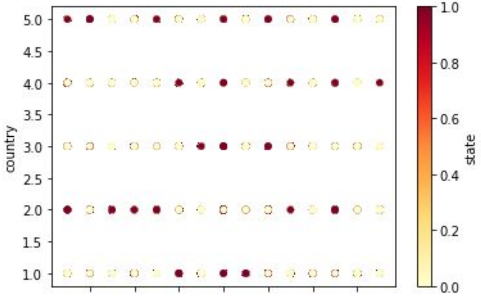
- Performed logistic regression analysis with the columns “goal rate”, “country”, “main_category”, and “backers”
- With this model we tested 74669 data points and predicted that 24.8% of projects would succeed and that 75.2% of them would fail
- Found that the accuracy score of this model was 83%
- Coefficient of features
  - Goal rate = 0.621
  - Country = -0.113
  - Main Category = -0.034
  - Backer = 0.016


Logistic Regression Model
X1
X2
X3
θ_1
θ_2
θ_3
Happy
Sad
Inputs: X1,X2,X3 || Weights: Θ1,Θ2,Θ3 || Outputs: Happy or Sad
@dataaspirant.com

KNN Analysis Using Main Category and Country

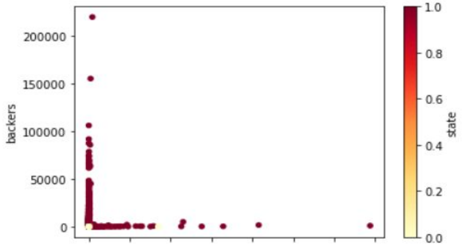
- Performed KNN analysis using “main_category” and “country”
  - Paired these two features together before they showed negative coefficients in our Logistic Regression test
- Graphed the relationship between main_category and country
- Used the elbow method to find k=3 due to the insignificant change after running the algorithm for k=3, k=5, and k=20
- With this model we predicted 15050 successes and 59619 failures





KNN Analysis Using Goal Rate and Backers

- Performed KNN analysis with the columns that we assume have the most useful features as shown in our coefficient with Logistic Regression: “Goal rate” and “Backers”
- Used Elbow method to find the best K value
- K = 5, result in 36.2% classified as successful and 63.8% classified as failed
  - Accuracy with K = 5, is 98% correctly predicted
  - Since the accuracy is high enough we decide to use K = 5 as the best K value
  - Result confirm that “Goal rate” and “backers” are the best features



Questions

1. What coefficients did we find that have a positive relationship?
2. What coefficients did we find that have a negative relationship?
3. What was the best K value we found for KNN analysis using main_category and Country?

Thank You!