IDH30305

- Professor Srikanth Namuduri
- Mekha Abraham, Leonardo Rodriguez, Madheline Almanzar

KICSTARTER PREDICTION MODEL

Introduction

Kickstarter “help bring creative projects to life”. It is an American corporation devoted to crowdfunding projects worldwide through their platform. Kickstarter was funded in 2009 and have been bringing creative projects and startups to life ever since. Their business model is sustainable thanks to general public engaging in different projects, or, as they call it, “pledging”, in hopes that they will get something in return from the company. There is a rigorous process for creators to be featured on Kickstarter’s platform, which has built confidence among supporters (backers) to contribute monetarily (pledge) with such projects, making Kickstarter’s business model sustainable. There is a deadline established by which the startup must reach their monetary goal or, otherwise, backers won’t be charged for the money they pledged (get to keep their money).

Data

```
import csv
import numpy as np
import pandas as pd
from google.colab import drive
```

```
#In this step, we accessed our data on GoogleDrive. This was a way to import the dat
drive.mount('/content/drive')
```

```
Go to this URL in a browser: https://accounts.google.com/o/oauth2/auth?client_

Enter your authorization code:
··········
Mounted at /content/drive
```

```
#This funtion and the lines following allowed us to open and read the data set.
import os
```

```
os.chdir('/content/drive/My Drive/')
```

```
os.getcwd()
```

```
'/content/drive/My Drive'
```

```
#Here we defined the the dataset as df in a dataframe.
df = pd.read_csv("Data2.csv")
```

```
#Head allows us to see first 5 rows. Each column represents an important factor of a
df.head()
```

	ID	name	category	main_category	currency	deadline	goal
0	1000002330	The Songs of Adelaide & Abullah	Poetry	Publishing	GBP	10/9/2015 11:36	1000.0
1	1000004038	Where is Hank?	Narrative Film	Film & Video	USD	2/26/2013 0:20	45000.0
2	1000007540	ToshiCapital Rekordz Needs Help to Complete Album	Music	Music	USD	4/16/2012 4:24	5000.0
3	1000011046	Community Film Project: The Art of Neighborhoo...	Film & Video	Film & Video	USD	8/29/2015 1:00	19500.0
4	1000014025	Monarch	Restaurants	Food	USD	4/1/2016	50000.0

```
#Df.info gave us infor about the columns and the data type in each column. Some colu
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 999 entries, 0 to 998
Data columns (total 13 columns):
ID                999 non-null int64
name              999 non-null object
category          999 non-null object
main_category     999 non-null object
currency          999 non-null object
deadline          999 non-null object
goal              999 non-null float64
launched          999 non-null object
pledged           999 non-null float64
state             999 non-null object
backers           999 non-null int64
country           999 non-null object
usd pledged       986 non-null float64
dtypes: float64(3), int64(2), object(8)
memory usage: 101.5+ KB
```

Data Cleaning

```
#missing_values are defnied as 'n/a', '--', 'na'
missing_values = ["n/a", "--", "na"]


# Present working directory
pwd
```

```
'/content/drive/My Drive'
```

```
#We redefined the df again to exclude the missing values.
df = pd.read_csv("Data2.csv")


df = pd.read_csv("Data2.csv", na_values = missing_values)


#Full dataset
df
```

		Ya'll come an ...				2:29
982	1005787932	Hotdog Yogatote	Product Design	Design	USD	4/2/2014 21:39
983	1005790478	Travel Far: A... Guide to the Out-of-Body Expe...	Nonfiction	Publishing	USD	4/12/2015 0:00
984	1005800160	Who is Nellie Bly?	Painting	Art	USD	8/5/2012 5:17
985	100580052	Get It On The Shelf!	Fiction	Publishing	USD	7/3/2011 1:30
986	1005808602	Help Us Record Our Debut EP "Mayday"	Music	Music	USD	6/23/2011 2:03
987	1005820080	The Million Pound Shirt	Fashion	Fashion	GBP	11/25/2015 22:00

```
#To see how many missing values (NaN) we have in each column.
df.isnull().sum()
```

```
ID                0
name              0
category          0
main_category     0
currency          0
deadline          0
goal              0
launched          0
pledged           0
state             0
backers           0
country           0
usd pledged      13
dtype: int64
```

short film 2:43

```
df.columns
```

```
Index(['ID ', 'name ', 'category ', 'main_category ', 'currency ', 'deadline '
       'goal ', 'launched ', 'pledged ', 'state ', 'backers ', 'country ',
       'usd pledged '],
      dtype='object')
```

6.00

```
from pandas import Series, DataFrame
```

KICKSTARTER Theaters 20:08

```
df.isnull().sum()
```

```
ID                0
name              0
category          0
main_category     0
currency          0
deadline          0
goal              0
launched          0
pledged           0
```

```
df.dropna()
```

						18:06
27	1000117861	Ledr workbook: one tough journal!	Product Design	Design	USD	10/8/2016 2:00
28	1000120151	Feather Cast Furled Fly Fishing Leaders	Product Design	Design	AUD	8/22/2015 3:09
29	1000120287	BB130A	Public Art	Art	USD	3/24/2013 0:07
...	...	...	...	...	...	...
		GBS Detroit				

```
df.isnull().sum()
```

```
ID                0
name              0
category          0
main_category     0
currency          0
deadline          0
goal              0
launched          0
pledged           0
state             0
backers           0
country           0
usd pledged      13
dtype: int64
```

		Darker Shade of				3/31/2014

```
df.dropna(how='any')
```

	ID	name	category	main_category	currency	deadline
0	1000002330	The Songs of Adelaide & Abullah	Poetry	Publishing	GBP	10/9/2015 11:36
1	1000004038	Where is Hank?	Narrative Film	Film & Video	USD	2/26/2013 0:20
2	1000007540	ToshiCapital Rekordz Needs Help to Complete Album	Music	Music	USD	4/16/2012 4:24
3	1000011046	Community Film Project: The Art of Neighborhoo...	Film & Video	Film & Video	USD	8/29/2015 1:00
4	1000014025	Monarch Espresso Bar	Restaurants	Food	USD	4/1/2016 13:38
5	1000023410	Support Solar Roasted Coffee & Green Energy! ...	Food	Food	USD	12/21/2014 18:30
6	1000030581	Chaser Strips. Our Strips make Shots their B*tch!	Drinks	Food	USD	3/17/2016 19:05
7	1000034518	SPIN - Premium Retractable In-Ear Headphones w...	Product Design	Design	USD	5/29/2014 18:14
		STUDIO IN THE SKY - A				8/10/2014

```
df.isnull().sum()
```

```
ID               0
name             0
category         0
main_category    0
currency         0
deadline         0
goal             0
launched         0
pledged          0
state            0
backers          0
country          0
usd pledged     13
dtype: int64
```

13	1000064368	Survival Rings	Design	Design	USD	2/28/2015

```
#To see an specific row
df.iloc[150]
```

```
ID                                1000694855
name              STREETFIGHTERZ WHEELIE MURICA
category                           Film & Video
main_category                      Film & Video
currency                                    USD
deadline                        9/20/2014 6:59
goal                                      6500
launched                        8/6/2014 21:28
pledged                                    555
state                                undefined
backers                                      0
country                                    N,"0
usd pledged                                 NaN
Name: 150, dtype: object
```

```
#To drop rows with any NaN value
df.dropna(axis=0, how='any', inplace = True)
```

```
df.isnull().sum()
```

```
ID                0
name              0
category          0
main_category     0
currency          0
deadline          0
goal              0
launched          0
pledged           0
state             0
backers           0
country           0
usd pledged       0
dtype: int64
```

```
df.columns = ['ID','name', 'category', 'main_category', 'currency','deadline','goal'
```

```
#number of unique values in all columns
print(df.nunique())
```

```
ID                986
name              986
category          124
main_category      15
currency           12
deadline          985
goal              192
launched          986
pledged           675
state               5
backers           234
country            19
usd pledged       710
dtype: int64
```

```
#Distribution of data across state
percent_dist = round(df["state"].value_counts() / len(df["state"]) * 100,2)

print("State Percent: ")
print(percent_dist)

# Filtering only for successful and failed projects# Filte
kick_projects = df[(df['state'] == 'failed') | (df['state'] == 'successful')]
#converting 'successful' state to 1 and failed to 0
kick_projects['state'] = (kick_projects['state'] =='successful').astype(int)
print(kick_projects.shape)
```

```
State Percent:
failed        50.81
successful    38.24
canceled       8.82
live           1.83
suspended      0.30
Name: state, dtype: float64
(878, 13)
/usr/local/lib/python3.6/dist-packages/ipykernel_launcher.py:9: SettingWithCop
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: http://pandas.pydata.org/pandas-docs/sta
  if __name__ == '__main__':
```

```
#checking distribution of projects across various main categories
kick_projects.groupby(['main_category','state']).size()
```

```
main_category  state
Art            0          46
               1          40
Comics         0          14
```

```
# This line of code adds a column to the dataframe. The column is a true or false co
df['TF']= df['usd pledged']>= df['goal']
```

```
df.head()
```

	ID	name	category	main_category	currency	deadline	goal
0	1000002330	The Songs of Adelaide & Abullah	Poetry	Publishing	GBP	10/9/2015 11:36	1000.0
1	1000004038	Where is Hank?	Narrative Film	Film & Video	USD	2/26/2013 0:20	45000.0
2	1000007540	ToshiCapital Rekordz Needs Help to Complete Album	Music	Music	USD	4/16/2012 4:24	5000.0
3	1000011046	Community Film Project: The Art of Neighborhoo...	Film & Video	Film & Video	USD	8/29/2015 1:00	19500.0
4	1000014025	Monarch Espresso Bar	Restaurants	Food	USD	4/1/2016 13:38	50000.0

```
maincategory = pd.get_dummies(df['main_category'])
```

```
maincategory
```

	Art	Comics	Crafts	Dance	Design	Fashion	Film & Video	Food	Games	Journalisr
0	0	0	0	0	0	0	0	0	0	(
1	0	0	0	0	0	0	1	0	0	(
2	0	0	0	0	0	0	0	0	0	(
3	0	0	0	0	0	0	1	0	0	(
4	0	0	0	0	0	0	0	1	0	(
5	0	0	0	0	0	0	0	1	0	(
6	0	0	0	0	0	0	0	1	0	(
7	0	0	0	0	1	0	0	0	0	(
8	0	0	0	0	0	0	1	0	0	(
9	0	0	0	0	0	0	0	0	0	(
10	0	0	0	0	0	0	0	0	0	(
11	0	0	1	0	0	0	0	0	0	(
12	0	0	0	0	0	0	0	0	1	(
13	0	0	0	0	1	0	0	0	0	(
14	0	1	0	0	0	0	0	0	0	(
15	0	0	0	0	0	0	0	0	0	(
16	0	0	0	0	0	0	0	0	0	(
17	0	0	0	0	0	0	0	1	0	(
18	0	0	0	0	0	1	0	0	0	(
19	0	0	0	0	0	0	0	0	0	(
20	0	0	0	0	0	0	0	1	0	(

```
#We replaced the column main_category with integers so we can run logistical regress
>>> df.replace(['Art', 'Comics','Crafts','Dance', 'Design', 'Fashion', 'Film & Video
```

	ID	name	category	main_category	currency	deadline
0	1000002330	The Songs of Adelaide & Abullah	Poetry	13	GBP	10/9/2015 11:36
1	1000004038	Where is Hank?	Narrative Film	7	USD	2/26/2013 0:20
2	1000007540	ToshiCapital Rekordz Needs Help to Complete Album	11	11	USD	4/16/2012 4:24
3	1000011046	Community Film Project: The Art of Neighborhoo...	7	7	USD	8/29/2015 1:00
4	1000014025	Monarch Espresso Bar	Restaurants	8	USD	4/1/2016 13:38
5	1000023410	Support Solar Roasted Coffee & Green Energy! ...	8	8	USD	12/21/2014 18:30
6	1000030581	Chaser Strips. Our Strips make Shots their B*tch!	Drinks	8	USD	3/17/2016 19:05
7	1000034518	SPIN - Premium Retractable In-Ear Headphones w...	Product Design	5	USD	5/29/2014 18:14
8	100004195	STUDIO IN THE SKY - A Documentary Feature Film...	Documentary	7	USD	8/10/2014 21:55
9	100004721	Of Jesus and Madmen	Nonfiction	13	CAD	10/9/2013 18:19
10	100005484	Lisa Lim New CD!	Indie Rock	11	USD	4/8/2013 6:42
11	1000055792	The Cottage Market	3	3	USD	10/2/2014 17:11
		G-Spot Place for Gamers to				3/25/2016

Logistic Regression

In this section, three different machine learning models for classification will be applied to the data, in order to create a model to classify projects into successes and failures.

```
X = df[['main_category', 'goal', 'backers']]
```

```
X.head()
```

	main_category	goal	backers
0	Publishing	1000.0	0
1	Film & Video	45000.0	3
2	Music	5000.0	1
3	Film & Video	19500.0	14
4	Food	50000.0	224

```
x = X.replace(['Art', 'Comics','Crafts','Dance', 'Design', 'Fashion', 'Film & Video'
```

```
#x is our redefined adjusted data set that only includes numbers and the three x-var
x
```

	main_category	goal	backers
0	13	1000.0	0
1	7	45000.0	3
2	11	5000.0	1
3	7	19500.0	14
4	8	50000.0	224
5	8	1000.0	16
6	8	25000.0	40
7	5	125000.0	58
8	7	65000.0	43
9	13	2500.0	0
10	11	12500.0	100
11	3	5000.0	0
12	9	200000.0	0

```
#We created a numpy arrays to facilitate regression.
x = np.array(x)
x[:3]
```

```
array([[1.3e+01, 1.0e+03, 0.0e+00],
       [7.0e+00, 4.5e+04, 3.0e+00],
       [1.1e+01, 5.0e+03, 1.0e+00]])
```

```
#Our Y value is the sucess of the kickstarter which we had defined earlier using the
y = df[['TF']]
```

20	8	500.0	3

```
y.head()
```

	TF
0	False
1	False
2	False
3	False
4	True

```
y = np.array(y)
y[:3]
```

```
array([[False],
       [False],
       [False]])
```

```
from sklearn.linear_model import LogisticRegression
```

```
from sklearn.model_selection import train_test_split
```

```
# This logistc regression.
x_train, x_test, y_train, y_test = train_test_split(x,y, test_size=0.2, random_state
```

```
logreg = LogisticRegression()
logreg.fit(x_train,y_train)
```

```
/usr/local/lib/python3.6/dist-packages/sklearn/linear_model/logistic.py:433: F
  FutureWarning)
/usr/local/lib/python3.6/dist-packages/sklearn/utils/validation.py:761: DataCo
  y = column_or_1d(y, warn=True)
LogisticRegression(C=1.0, class_weight=None, dual=False, fit_intercept=True,
          intercept_scaling=1, max_iter=100, multi_class='warn',
          n_jobs=None, penalty='l2', random_state=None, solver='warn',
          tol=0.0001, verbose=0, warm_start=False)
```

```
# Making predictions
y_hat_train = logreg.predict(x_train)
```

```
y_hat_test = logreg.predict(x_test)
```

```
print("Logistic regression score for training set:", round(logreg.score(x_train, y_t
print("Logistic regression score for test set:", round(logreg.score(x_test, y_test),
```

```
Logistic regression score for training set: 0.8566
Logistic regression score for test set: 0.83333
```