What makes a Crowdfunding Successful?

...

Lorraine Zhou Sahil Natu Bret Jaco Jonathan Jan

August 9, 2021

01 Problem

02 Dataset

03 Models

04 Conclusions

Problem Overview

- What factors contribute to a successful crowdfunding campaign?
  - Often, projects fail to meet their funding goal or are cancelled due to management issues

- Global Crowdfunding platform for entrepreneurs
- Supports millions of investors and creators

Project objective:

Build a classification model to predict crowdfunding success and get insights from the dataset

Data Source and Cleaning

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 334474 entries, 0 to 378660
Data columns (total 15 columns):
 #   Column            Non-Null Count   Dtype
---  ------            --------------   -----
 0   ID                334474 non-null  int64
 1   name              334474 non-null  object
 2   category          334474 non-null  object
 3   main_category     334474 non-null  object
 4   currency          334474 non-null  object
 5   deadline          334474 non-null  datetime64[ns]
 6   goal              334474 non-null  float64
 7   launched          334474 non-null  datetime64[ns]
 8   pledged           334474 non-null  float64
 9   state             334474 non-null  object
 10  backers           334474 non-null  int64
 11  country           334474 non-null  object
 12  usd pledged       334263 non-null  float64
 13  usd_pledged_real  334474 non-null  float64
 14  usd_goal_real     334474 non-null  float64
dtypes: datetime64[ns](2), float64(5), int64(2), object(6)
memory usage: 40.8+ MB
```

Some columns are numeric, and others are categorical.

1.The missing, duplicates value processing

2.Drop the index that state is **'undefined', 'suspended', 'canceled'.**

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 334474 entries, 0 to 378660
Data columns (total 7 columns):
 #   Column            Non-Null Count   Dtype
---  ------            --------------   -----
 0   main_category     334474 non-null  object
 1   state             334474 non-null  object
 2   backers           334474 non-null  int64
 3   country           334474 non-null  object
 4   usd_pledged_real  334474 non-null  float64
 5   usd_goal_real     334474 non-null  float64
 6   launched_year     334474 non-null  int64
dtypes: float64(2), int64(2), object(3)
memory usage: 20.4+ MB
```

Drop the columns that is unnecessary data

Some columns will not be useful in our prediction models. These will be removed:

name	usd pledged	currency
category	deadline	launched
goal	pledged	id

4	main_category_Comics	334474	non-null	uint8
5	main_category_Crafts	334474	non-null	uint8
6	main_category_Dance	334474	non-null	uint8
7	main_category_Design	334474	non-null	uint8
8	main_category_Fashion	334474	non-null	uint8
9	main_category_Film & Video	334474	non-null	uint8
10	main_category_Food	334474	non-null	uint8
11	main_category_Games	334474	non-null	uint8
12	main_category_Journalism	334474	non-null	uint8
13	main_category_Music	334474	non-null	uint8
14	main_category_Photography	334474	non-null	uint8
15	main_category_Publishing	334474	non-null	uint8
16	main_category_Technology	334474	non-null	uint8
17	main_category_Theater	334474	non-null	uint8

19	country_BE	334474	non-null	uint8
20	country_CA	334474	non-null	uint8
21	country_CH	334474	non-null	uint8
22	country_DE	334474	non-null	uint8
23	country_DK	334474	non-null	uint8
24	country_ES	334474	non-null	uint8
25	country_FR	334474	non-null	uint8
26	country_GB	334474	non-null	uint8
27	country_HK	334474	non-null	uint8
28	country_IE	334474	non-null	uint8
29	country_IT	334474	non-null	uint8
30	country_JP	334474	non-null	uint8
31	country_LU	334474	non-null	uint8
32	country_MX	334474	non-null	uint8

Dummy variable processing

One hot encoding :

main category, country, usd_pledged_real

Label Encoding :

Change the status column values to :

successful: 1

failed: 0

Exploratory Data Analysis

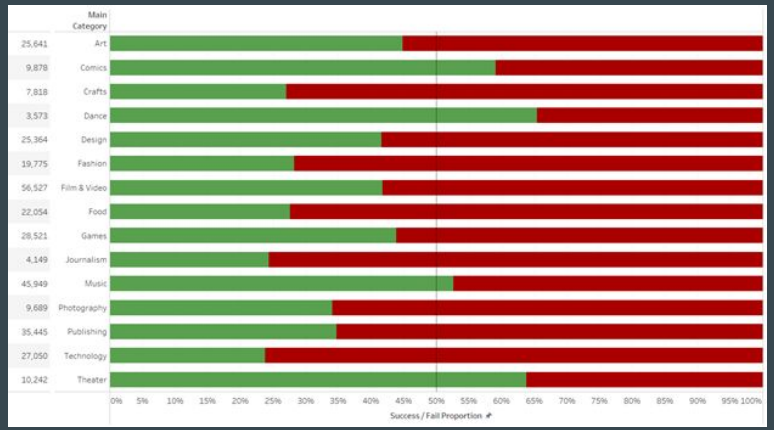
Exploratory Data Analysis

Why?

- Unearth patterns
- Understand the distribution of data
- Interpret the data better

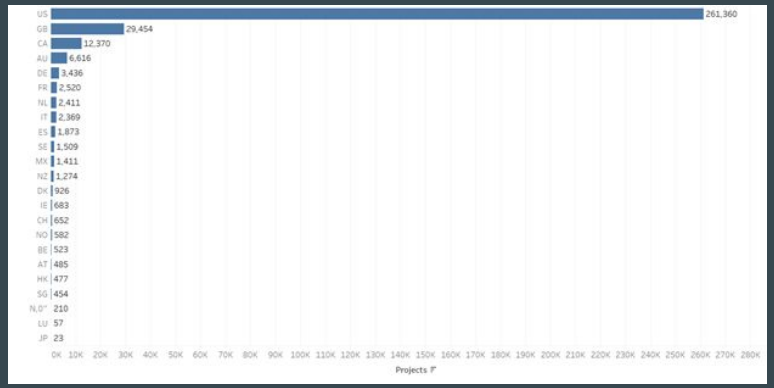
Dataset Features:

- 378,661 records
- 5 Numerical Variables
- 9 Categorical Variables



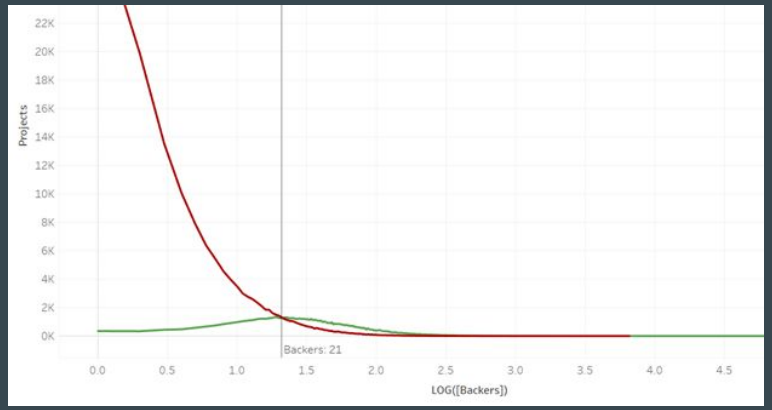
Success/Failure by Main Category

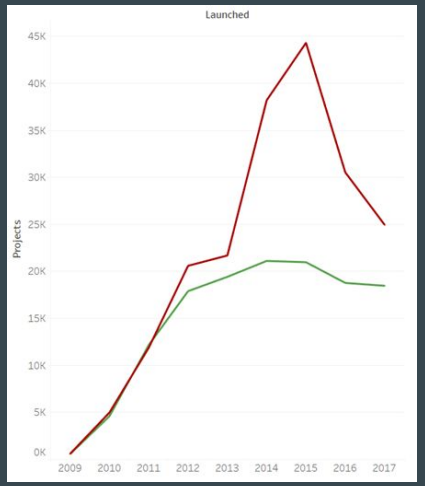
- Mean Success Rate: 40%
- High of 65% for Dance, Theater
- Low of 25% for Journalism, Tech



Projects by Country

- Maximum projects from the US, with UK being a distant second
- 78% of the projects from the US



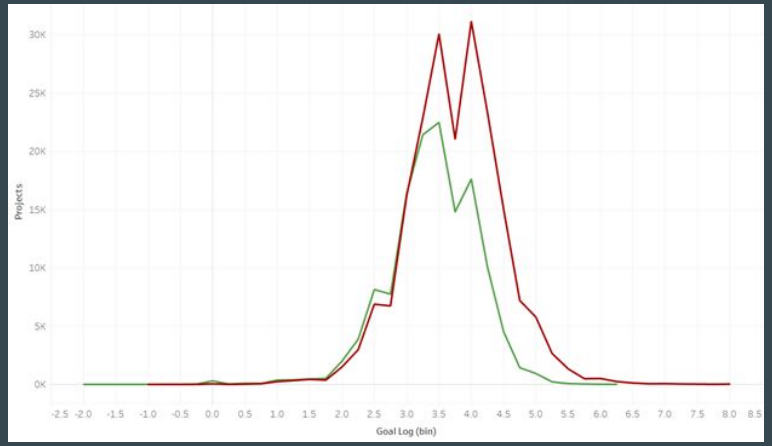


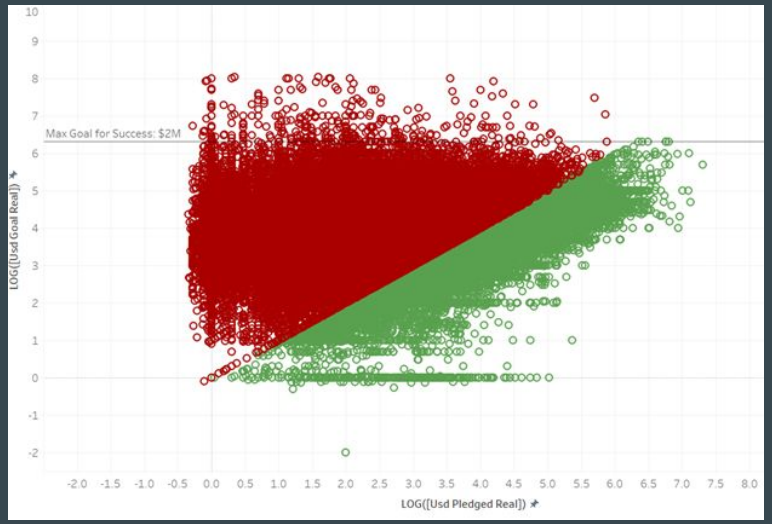
Success/Failure by No. of Backers

- More projects succeed than fail with increase in number of backers
- Threshold at 21 backers, beyond which likelihood of success is higher than failure

Success/Failure by Launch Year

- Rising number of projects every year
- Dip after 2015 as many projects would be still ongoing as of Jan 2018





Success/Failure by USD Goal Amount

- More projects fail than succeed as Goal amount increases
- Threshold at USD 1000, beyond which likelihood of failure is higher than success

Projects by USD Goal vs USD Pledged Amounts

- Many projects have secured funding more than goal amount
- No projects with goal higher than USD 2 million have succeeded

Predictive Models

Classifiers

- Logistic Regression
- Decision Tree
- Random Forest
- K-Nearest Neighbors

Predictors

- Main Category
- Number of Backers
- Country
- Pledged Amount in USD
- Goal Amount in USD
- Year Launched

Target Variable

State (successful or failed?)

Train/Test Set Split

- Training Set: 70% of Data (234,132 rows)
- Testing Set: 30% of Data (100,342 rows)

Logistic Regression

Training Accuracy Rate: 89.13%

Testing Accuracy Rate: 88.96%

Key Variables:

- Backers
- usd_goal_real

Feature Importance

```
Logistic Regression Variable Importance:
Feature: backers, Score: 0.04803
Feature: usd_goal_real, Score: -0.00021
Feature: launched_year, Score: -0.00041
Feature: cat_Art, Score: 0.00007
Feature: cat_Comics, Score: -0.00001
Feature: cat_Crafts, Score: -0.00005
Feature: cat_Dance, Score: 0.00005
Feature: cat_Design, Score: -0.00006
Feature: cat_Fashion, Score: -0.00009
Feature: cat_Film & Video, Score: 0.00024
```

Confusion Matrix

```
Logistic Regression Classifier -
Confusion Matrix:
[[57780  2487]
 [ 8594 31482]]
```

Decision Tree

Training Accuracy Rate: 98.80% (Overfitting!)

Testing Accuracy Rate: 90.01%

Key Variables:

- Backers
- usd_goal_real

Feature Importance

```
Decision Tree Variable Importance:
Feature: backers, Score: 0.70323
Feature: usd_goal_real, Score: 0.19087
Feature: launched_year, Score: 0.04095
Feature: cat_Art, Score: 0.00298
Feature: cat_Comics, Score: 0.00303
Feature: cat_Crafts, Score: 0.00148
Feature: cat_Dance, Score: 0.00124
Feature: cat_Design, Score: 0.00418
Feature: cat_Fashion, Score: 0.00266
Feature: cat_Film & Video, Score: 0.00522
```

Confusion Matrix

```
Decision Tree Classifier -
Confusion Matrix:
[[55611  4656]
 [ 5364 34712]]
```

Random Forest

Training Accuracy Rate: 98.30% (Overfitting!)

Testing Accuracy Rate: 91.81%

Key Variables:

- Backers
- usd_goal_real

Feature Importance

```
Random Forest Variable Importance:
Feature: backers, Score: 0.70257
Feature: usd_goal_real, Score: 0.20542
Feature: launched_year, Score: 0.03480
Feature: cat_Art, Score: 0.00224
Feature: cat_Comics, Score: 0.00290
Feature: cat_Crafts, Score: 0.00115
Feature: cat_Dance, Score: 0.00144
Feature: cat_Design, Score: 0.00383
Feature: cat_Fashion, Score: 0.00204
Feature: cat_Film & Video, Score: 0.00318
```

Confusion Matrix

```
Random Forest Classifier -
Confusion Matrix:
[[56341  3926]
 [ 4291 35785]]
```

K-Nearest Neighbors (with CV)

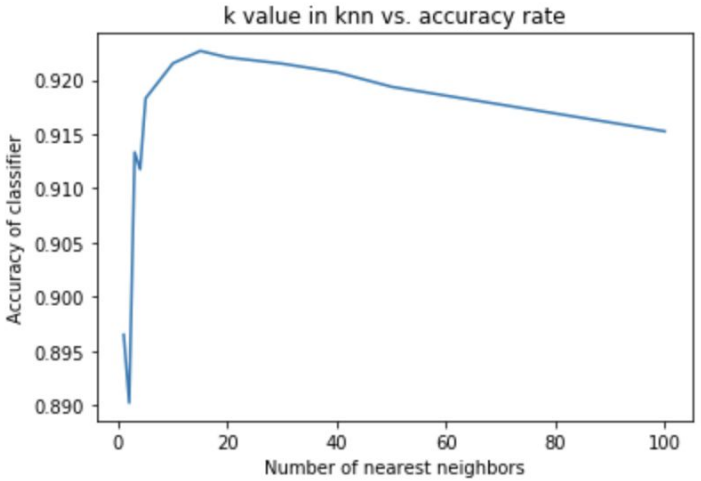
When k = 17:

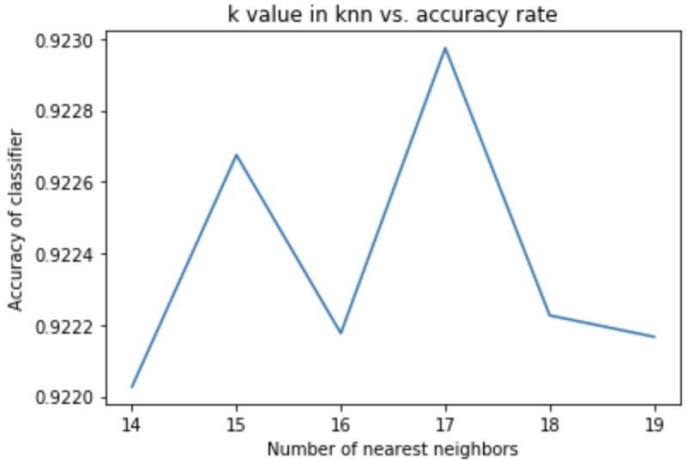
Training Accuracy Rate: 92.87%

Testing Accuracy Rate: 92.30% (Best so far!)

Confusion Matrix

```
KNN Confusion Matrix (K = 17):
[[56217  4050]
 [ 3679 36397]]
```





Turns Out the Best Model is.....

KNN Model with k = 17!

(92.30%)

And Most Important Variables are.....

Number of Backers

Goal Amount in USD

Conclusions

- Project categories influence success: Dance and Theater projects had a 65% success rate, while Technology and Journalism only 25%

- As expected, more backers increases success rate and a higher goal amount decreases success rate

THANKS

Any questions?