DMA Fall ‘21:

Amanda Fields

Albert Vasquez

Mei Johnson

Weiao Lin

Tyler Freund

KICKSTARTER

Investigating Crowdfunding Campaign Success

CONTEXT

from the ***Imperfect Utopia*** campaign, a photobook documenting the remote islands of Japan

$74 billion

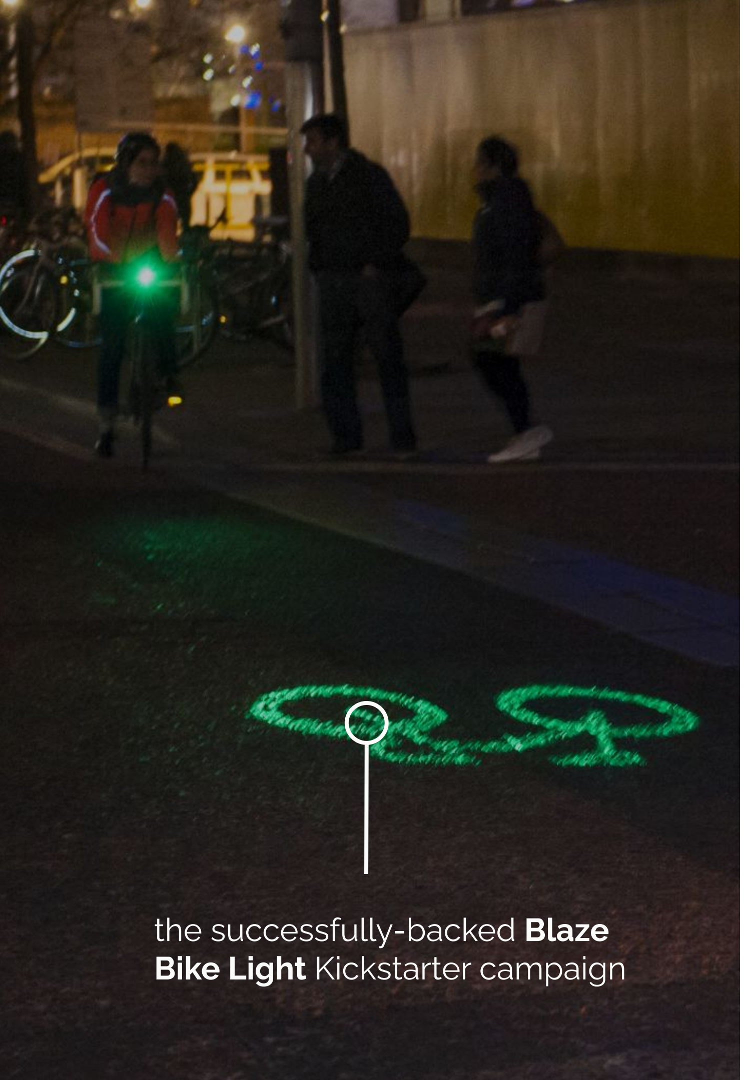
USD raised by **North American crowdfunding platforms** in 2020

(Source: Statista)

$777 million

USD raised by the crowdfunding platform **Kickstarter** in 2020

(Source: Kickstarter)



212,951

creative projects brought to life by Kickstarter

300,000+

part-time and full-time jobs created

(Source: Kickstarter, University of Pennsylvania)

My Coffee Box

Zen-Tek
Aquaponics

East Idaho
Aquarium Shirts

Mongol World
Art Tour

328,283

unsuccessfully-backed projects

What makes a Kickstarter campaign **successful**?

Can we **predict** a campaign's **success**?

DATA

from ***Reboot the Suit,*** the Smithsonian's campaign to restore Neil Armstrong's space suit

SOURCE

Kickstarter Projects Dataset
from Mickaël Mouillé on Kaggle

Included:

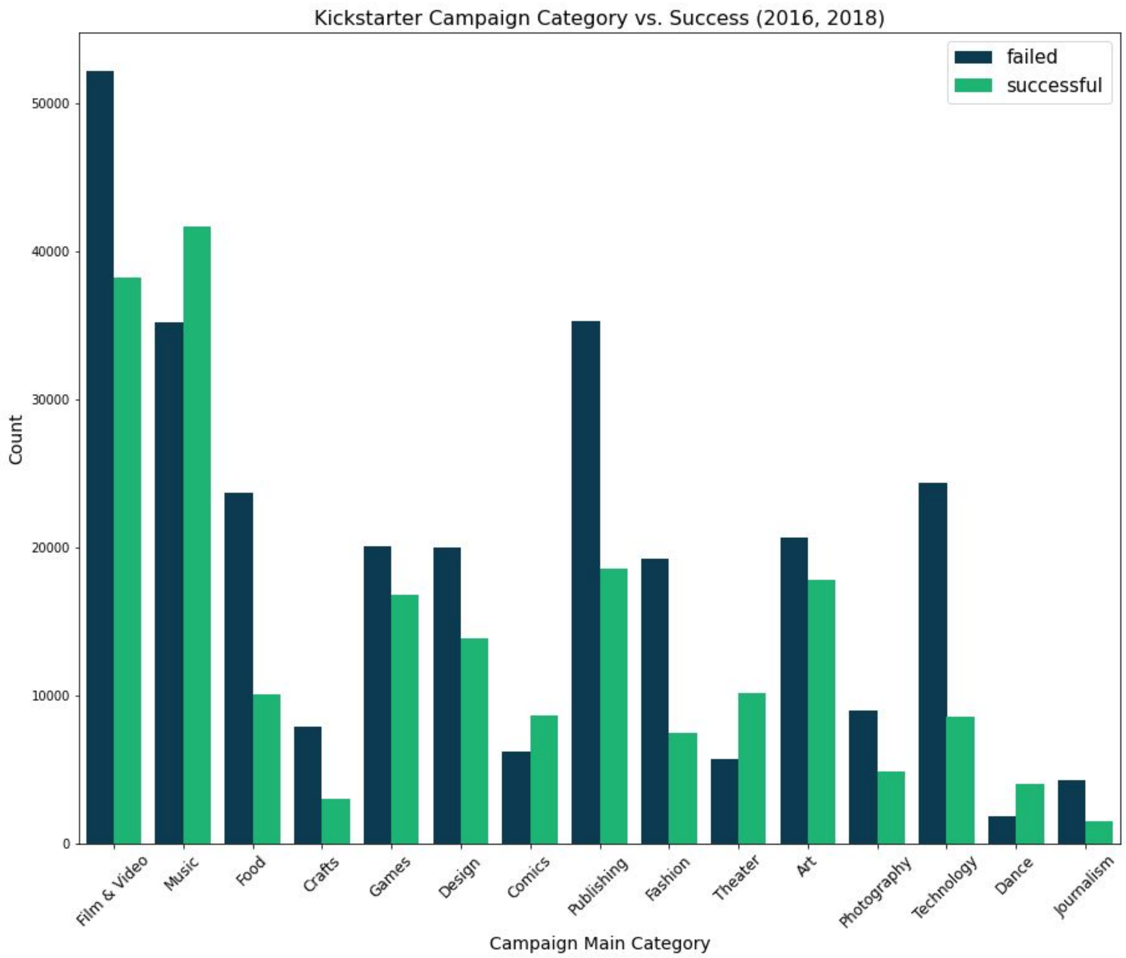
- ks-projects-201612.csv
  (data collected by Dec. 2016)
- ks-projects-201801.csv
  (data collected by Jan. 2018)

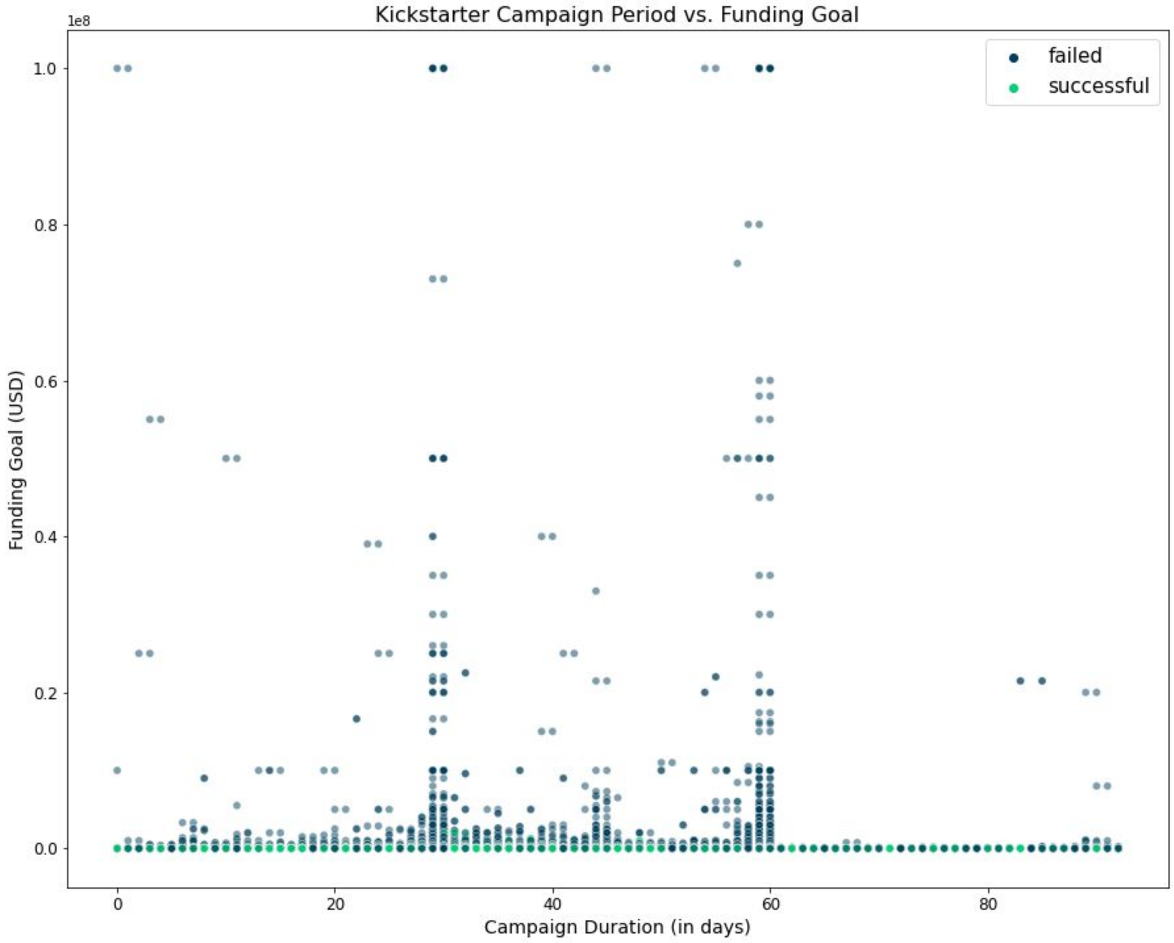
where each row = 1 project

ID *(campaign ID)*
name *(campaign name)*
main_category
category *(subcategory)*
currency
deadline
goal *(funding goal)*
launched
pledged
state *(state of campaign)*
backers
country
usdpledged
+ duration *(length, in days)*
+ percfunded

Category %

Film & Video: 18.41
Music: 15.66
Publishing: 10.97
Art: 7.82
Games: 7.51
Design: 6.90
Food: 6.88
Technology: 6.71
Fashion: 5.44
Theater: 3.23
Comics: 3.03
Photography: 2.83
Crafts: 2.22
Dance: 1.20
Journalism: 1.18



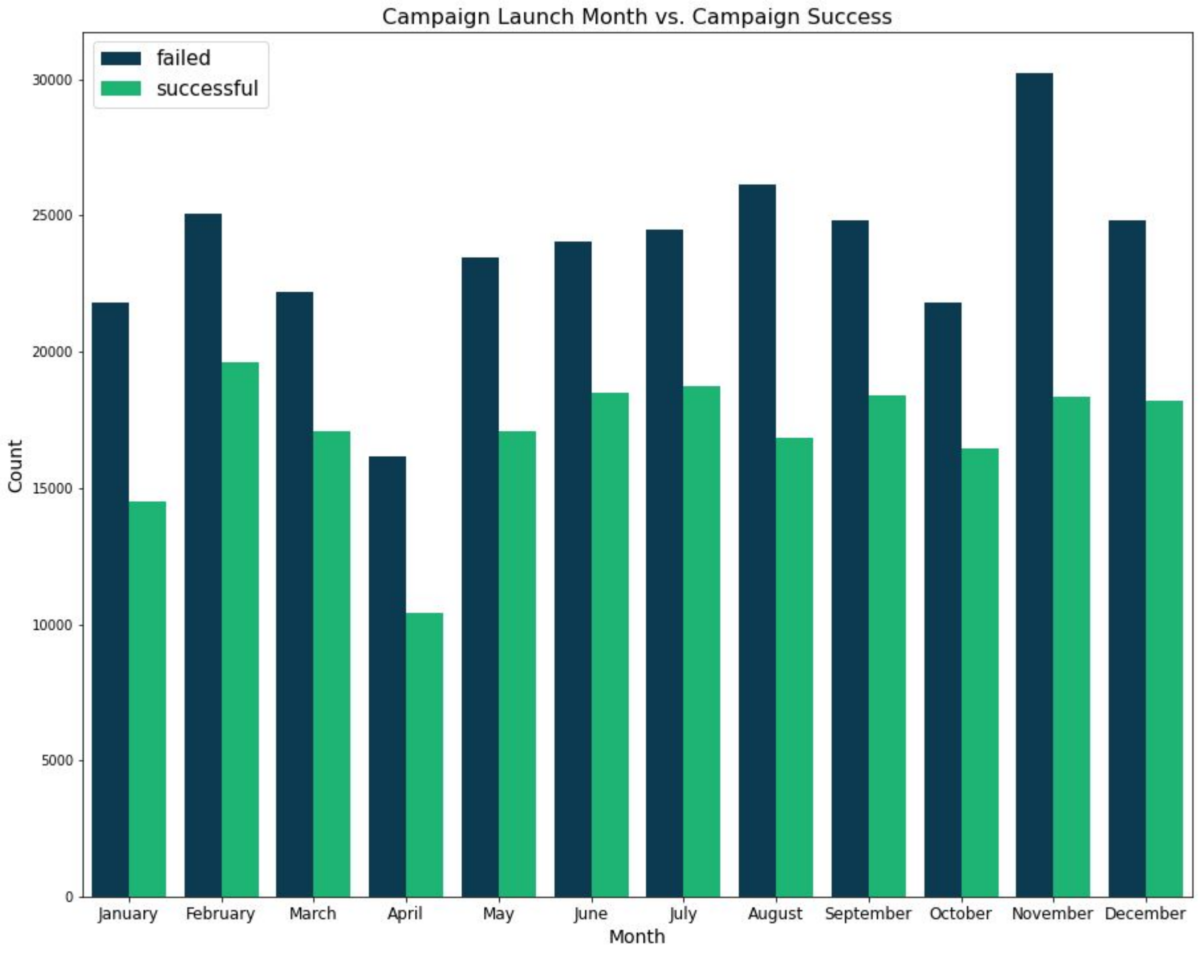


92 days

Longest Campaign

$ 100 mn

Largest Funding Goal

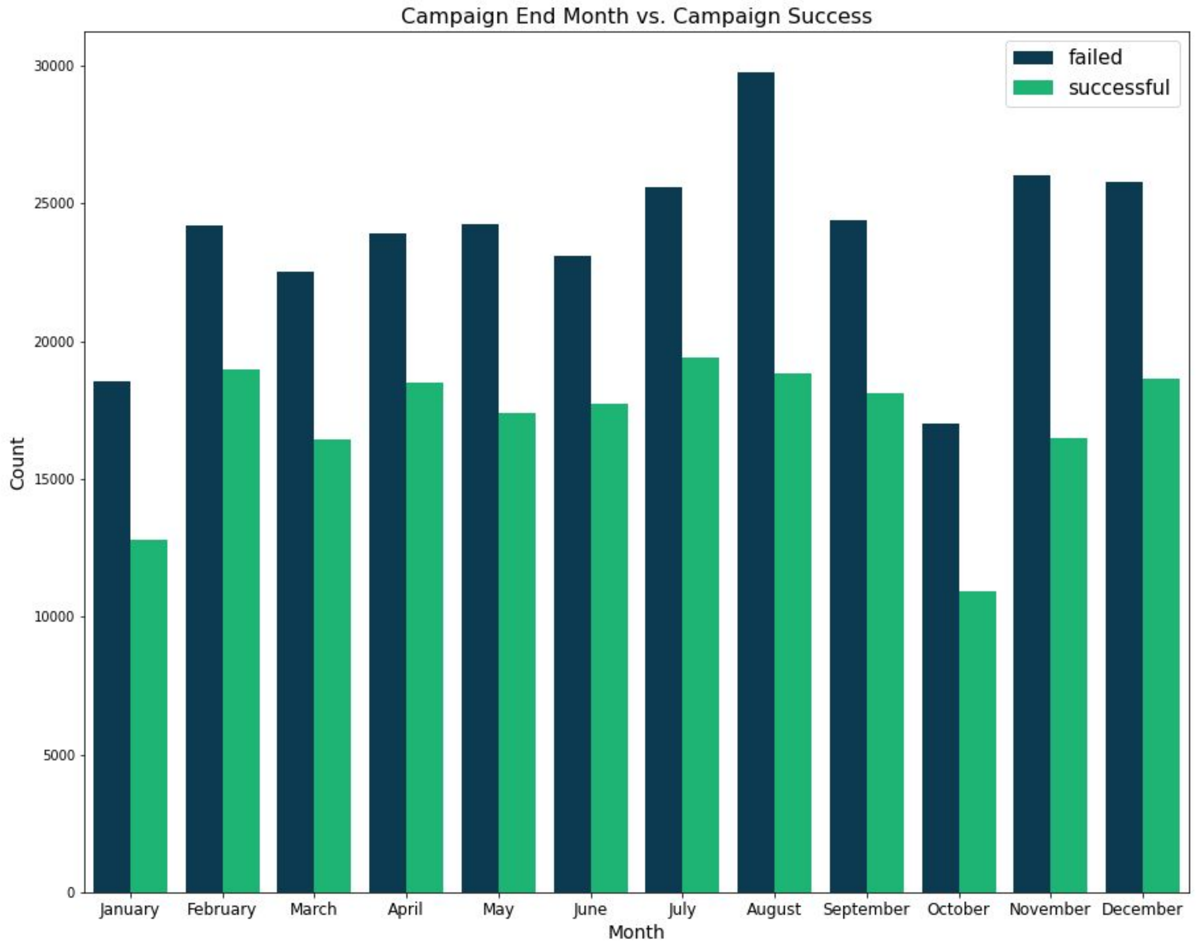


July

Most Popular Launch Month

Dec

Least Popular Launch Month



Aug

Most Popular End Month

Jan

Least Popular End Month

METHODS

from **The Narrows Public House**, a campaign to renovate an 1896 Nova Scotian public house

MODEL A1

:: Tyler

from **Barisier 2.0**, a campaign to create a coffee/tea-brewing alarm clock

CLEANING

- Removed NA/inf values
- Checked for negative values and duplicates
- Dropped instances where **`state`** == "live"

FEATURE ENGINEERING

- Added binary column **`outcome`**
    - 1 for success, else 0
- Log transformations of **`pledged`** & **`goal`**
- One-hot encoded **`categories`** feature
    - only kept the more prominent categories to avoid overfitting.
- Added **`month`** and **`year`** features
    - check if temporality is correlated with Kickstarter success.

MODELING

Used **scikit-learn** to implement the following::

- Random Forest
- Neural Network

FINDINGS

- Initial accuracy was low
  - in the mid-60% range for both models.

- After hyperparameter tuning:
  - **Random Forest:**
    - 87% training accuracy
    - 84% test accuracy
  - **Neural Network:**
    - 84% training accuracy
    - 83% test accuracy

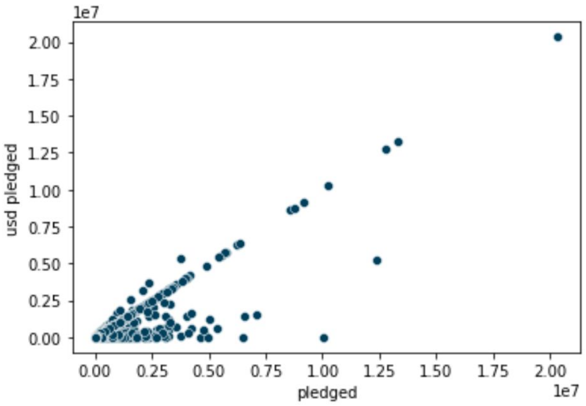
MODEL A2

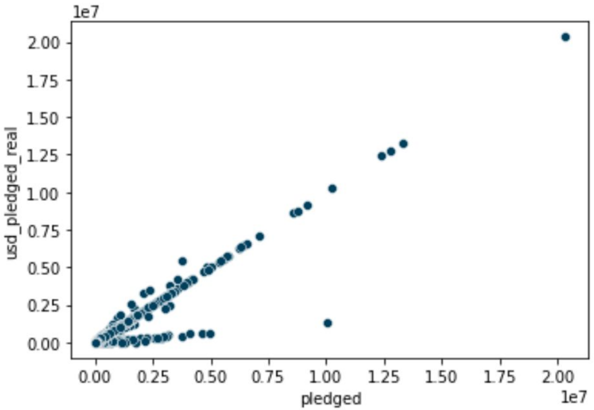
:: Weiao

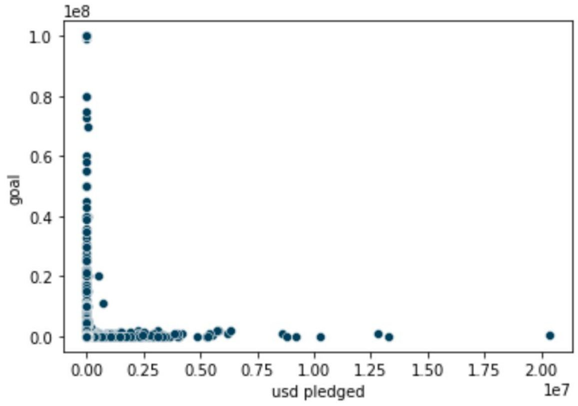
from **PLEISTOCENE PARK,** a documentary on recreating an Ice Age ecosystem in Siberia

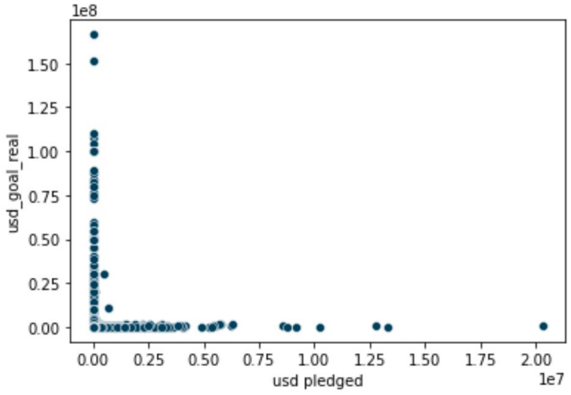
CLEANING & FEATURE ENGINEERING

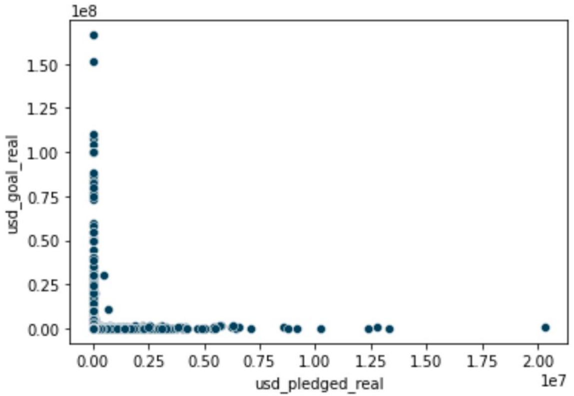
- Dropped all null values
- Collapsed values in `state`
  - 5 original values: “failed”, “successful”, “cancelled”, “live”, “suspended”
  - Removed “live”
  - Categorized “cancelled” and “suspended” into “failed”
- Dropped redundant columns
  - Visualized correlations between columns *(see next slide)*
  - Eliminated columns `usd_pledged_real` and `goal`
  - Kept columns `usd_pledged` and `usd_goal_real`
- Dropped other irrelevant columns
  - `ID`, `name`, `country`
- Streamlined time data
  - Eliminated years 1970, 2018 in `year`
  - Dropped `deadline` and kept `launched`
  - Split launched column into `year` & `month`











Correlations between selected columns of **ks18**, the 2018 Kickstarter data

MODELING

Used **scikit-learn** to implement the following::

- Decision Tree
- Random Forest

FINDINGS

- **Decision Tree:**
  - 97% training accuracy
  - 97% test accuracy
- **Random Forest:**
  - 99% training accuracy
  - 98% test accuracy

MODEL B

from **A Frog In the Fall**, a 300+ page comic book by Linnea Sterte

:: Amanda

Project We Love

CALIGULA hardcover book rare photos making of the cult film

It was the most expensive independent film in cinema history, and a debacle of equally epic...

by Thomas Negovan and 351 backers

Can a title predict a campaign's success?

What makes this better...

...than this?

Collinwood Needs Coffee!!!

Bright Coffee Bar aims to provide quality fresh, healthy gourmet foods and locally roasted coffe...

by Kimberly Homan

Funding unsuccessful

Project ended on July 28, 2014

CLEANING

- Dropped and manipulated NA values
  - Dropped `state` == "undefined"
  - Substituted values from `usd_pledged` into missing `usd_pledged_real` values
  - Substituted missing `country` values with country associated with value in `currency`

FEATURE ENGINEERING

- Created the following columns based on `name`:
  - `punctuation count`
  - `word count`
  - `character count`
  - `polarity`
    - Used Vader mean-sentiment-rating to determine sentiment of text

MODELING

Used **scikit-learn** to implement the following::

- Neural Network
- Logistic Regression Model

FINDINGS

- **Neural Network:**
  - ~60% training and testing accuracy
- **Logistic Regression Model:**
  - ~60% training and testing accuracy

CONCLUSION

from **The Tasman**, a campaign to
create a recycled Dutch oven & grill

IMPLICATIONS

- With the success of our models, we can provide **useful insight** to campaigners in regards to:
  - Timing of campaigns
  - Duration
  - Goal
  - Category
- We can also help investors predict the potential success of a project

FUTURE RESEARCH

- Delve deeper into **campaign categories**
  - Why are certain campaign categories more successful than others?
- Examine the **causal effect** of certain features on campaign performance
  - Campaign titles
  - Campaign images

APPENDIX

SOURCES

- Data courtesy of **Mickaël Mouillé** from Kaggle
- All images provided by **Kickstarter**, with the exception of the Appendix image, provided by **Death to Stock**
- Font: Raleway
- Deck Design: Mei

Tyler’s Workflow

Research question: What makes a kickstarter campaign successful? Can we build a model that predicts a successful campaign?

Cleaning:

- Removed na/inf values
- Checked for negative values and duplicates
- Dropped instances where state was “live” since we are focused on the outcome.

Feature engineering:

- Added binary column “outcome” (serving as my target variable) 1 for success, else 0.
- Log transformations of “pledged” and “goal ” features.
- One hot encoded “categories” feature and only kept the more prominent categories to avoid overfitting.
- Added “month” and “year” features in case temporality is correlated with kickstarter success.

Modeling:

- Trained a random forest model and a neural network (Both using SKlearn)

Findings:

- Initially my accuracy was rather low (in the mid 60% range) for both models.
- Tuning the hyperparameters of my model and adjusting my feature selection yielded much better results.
- Boosted random forest model up to 87% training accuracy and 84% test accuracy.
- Boosted Neural network up to 84% training accuracy and 83% test accuracy.

Weiao’s work

Research goal

- Build a model to predict a successful campaign in the 2018 Kickstart projects dataset

Data Processing & EDA

- Drop all null values in the dataset
- There are five different values in the “state” column, “failed”, “successful”, “cancelled”, “live”, “suspended”, I removed “live”, categorized “cancelled” and “suspended” into “failed” so we only have two values in the column, “failed” and “successful”
- There are many columns that seem to be redundant like “usd pledged” and “usd_pledged_real”, and “goal” and “usd_goal_real”, by visualizing at their correlations between each other, we can see that they are basically the same, so I can drop columns “usd_pledged_real” and “goal” from the dataframe and only keep columns “usd pledged” and “usd_goal_real”
- Drop other irrelevant columns “ID”, “name”, “country”
- For time data, drop “deadline” and keep “launched”, and break down “launched” column into columns “year” and “month”
- Eliminate years 1970 and 2018 in “year” column as these two years have too little counts

Modeling

- Decision tree: accuracy 97%
- Random forest: accuracy 99%

Amanda’s Work

-data cleaning (drop/manipulated null values)

-I was interested in whether or not we could predict a campaign’s success based on the title alone. Looking at number of characters, number of words, punctuation count, and sentiment analysis of the words.

- Used a neural model, and logistic regression model, both were in the %60 accuracy range.

Implications

Helping Existing and future campaigns

- With the success of our models, we can potentially provide useful insight to campaigners in regards to:
    - Timing of campaigns
    - Duration
    - Goal
    - Category

Future research areas:

- Delve deeper into campaign categories
    - Why are certain campaign categories more successful than others?
- Estimate the causal effect of certain features on campaign performance
    - Eg: Campaign titles