KICKSTARTER

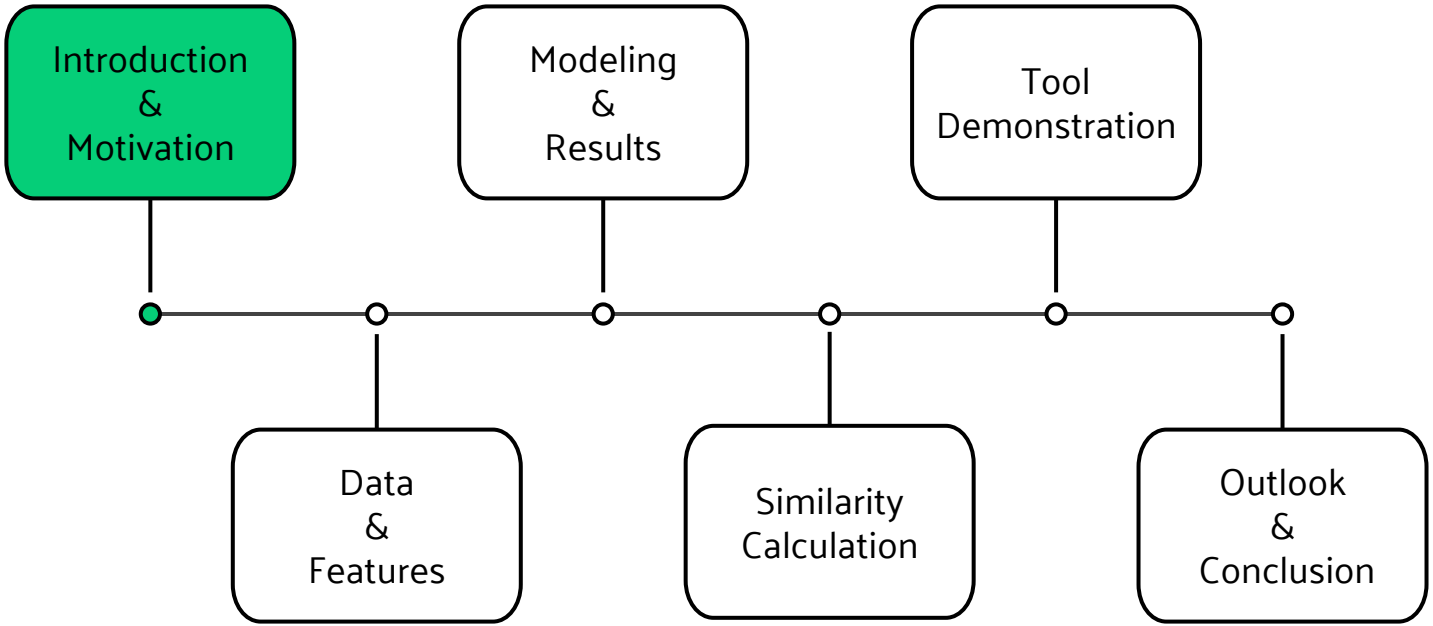
Crowdfunding Project Idea Validator

Presented by

Linxia Liu
Teng-Yun (Jacob) Chung
Vivian Kang
Yuan Liu
Zelong Qian

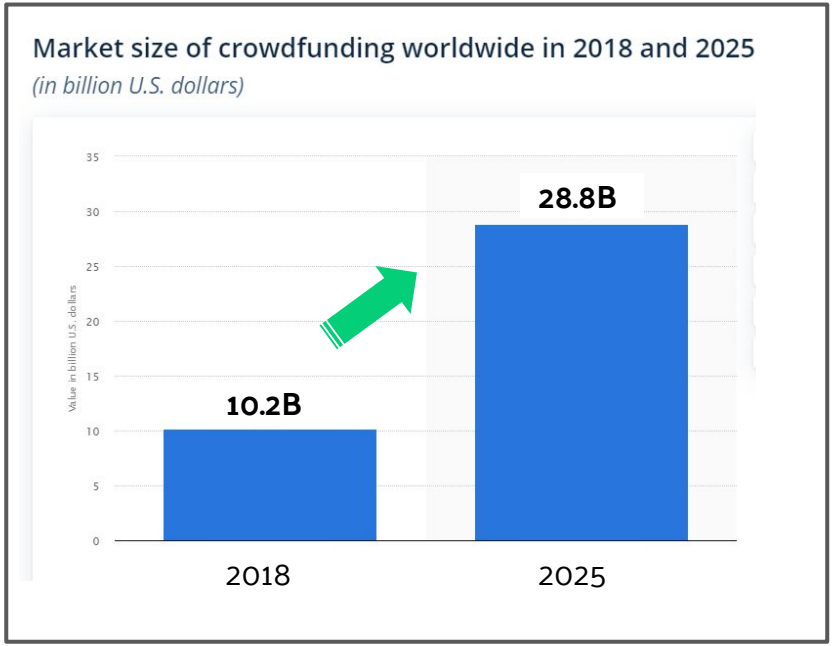
June 2020

Presentation Outline



Crowdfunding & Kickstarter

Crowdfunding: Rapidly Expanding Market



Kickstarter: Global Crowdfunding Platform

Projects and Dollars

Category	Launched Projects	Total Dollars	Successful Dollars	Unsuccessful Dollars
All	487,268	$5.02 B	$4.50 B	$471 M

Category	Successfully Funded Projects	Less than $1,000 Raised	$1,000 to $9,999 Raised	$10,000 to $19,999 Raised
All	182,655	23,840	98,769	26,239

Category	Unsuccessfully Funded Projects	0% Funded	1% to 20% Funded	21% to 40% Funded
All	301,601	55,598	194,289	30,320

37% success

Market: A tremendous potential for the demand for crowdfunding services

Need: A predictive analytics tool to improve the success rate of crowdfunding

Source: 1. Global crowdfunding market and forecast. https://www.statista.com/statistics/1078273/global-crowdfunding-market-size/; 2. Project statistics on Kickstarter. https://www.kickstarter.com/help/stats?ref=about_subnav

Project GOAL

KICKSTARTER

Develop a tool that could

- **Evaluate** people's ideas based **upon current Kickstarter projects**
- **Give** insightful feedback in the aspects of **potential success rate**
- **Provide** overall competitiveness with respect to **similar Kickstarter projects**

Input idea and relevant info

Description: e.g. 'I have a great idea...'

Category: games

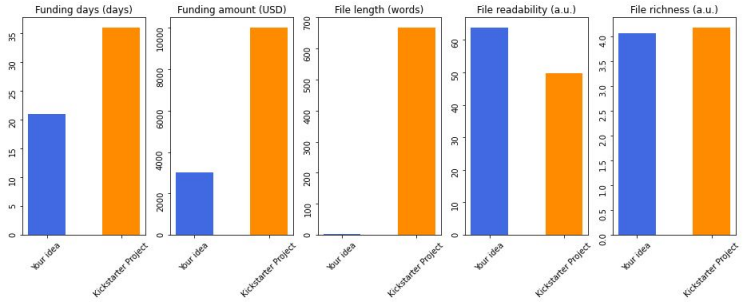
Amount: e.g. '5000'

Duration: e.g. '21'

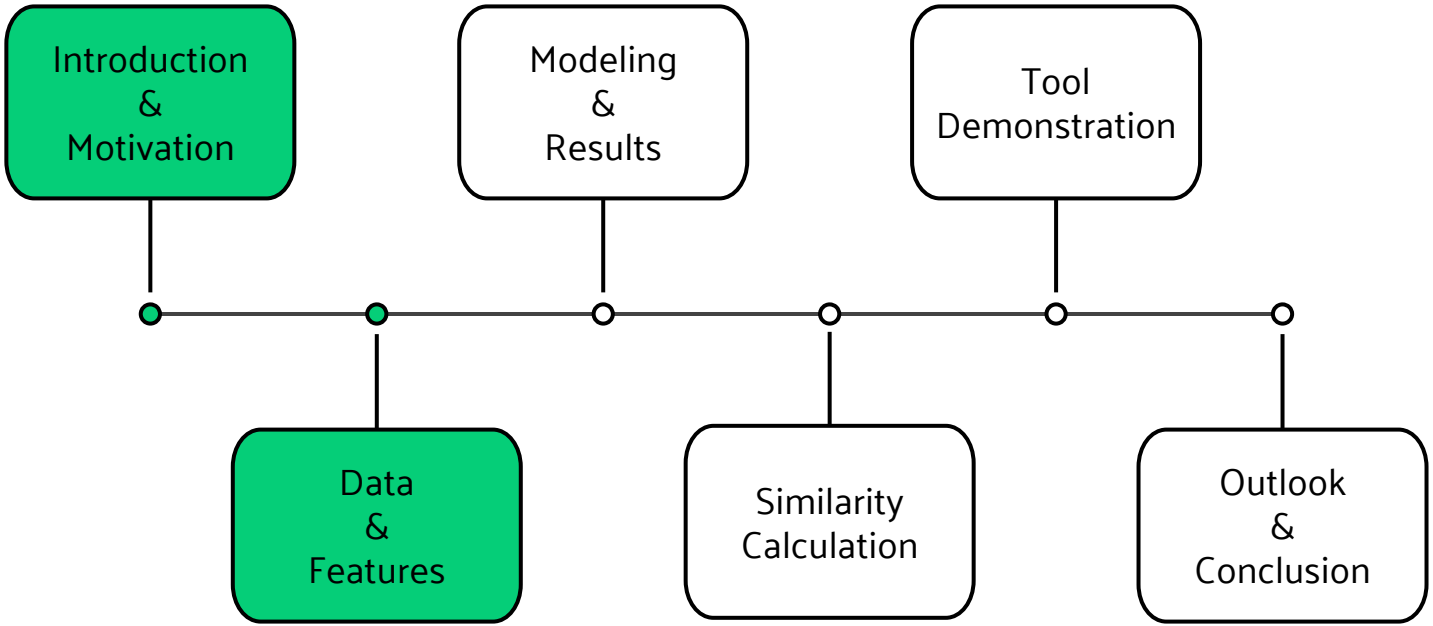
Submit

Generate predictive analysis results

```
==========
KICKSTARTER IDEA VALIDATOR RESULTS:
==========

Based on your input, the forecasted results are:
=> Predicted results: SUCCESSFUL
=> Successful probability: 54%
==========

Most similar projects on Kickstarter are identified for further comparison:
==========
NAME: 1ST & GOAL DARTS
DESCRIPTION: Double-sided dartboard (Football Dartboard & Standard Dartboard)

STATUS: FAILED
SIMILAR LEVEL: 55%
DETAILED COMPARISON:
```



Presentation Outline

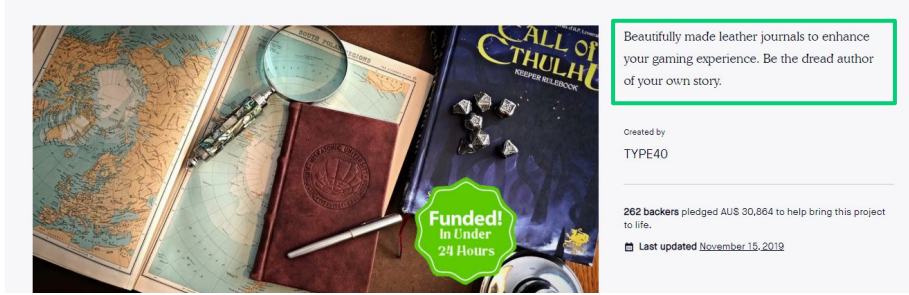


Data Collection

Original Kaggle Dataset Overview

FEATURE NAME	DATA FORMAT	# OF NA
id	int64	0
backers_count	int64	0
blurb	object	0
currency	object	0
goal	int64	0
launched_at	datetime64[ns]	0
deadline	datetime64[ns]	0
location.country	object	1
name	object	2
usd_pledged	float64	0
slug	object	0
spotlight	bool	0
staff_pick	bool	0
static_usd_rate	float64	0
state	object	0
year	int64	0
month	int64	0
day	int64	0
hour	int64	0
days_to_deadline	int64	0
goal_USD	float64	0
category_name	object	0
category_slug	object	0
blurb_length	int64	0
location_type	object	0
location_country	object	0
location_state	object	0
location_displayable_name	object	0
binary_state	object	0

"Successful", "Failed" 10058 records



- **Problem:** blurb does not provide sufficient text information for analysis
- **Solution:** web scraping the entire story section

Source: 2018-2019 Kickstarter project information collected from Kickstarter.com and https://www.kaggle.com/toshimelonhead/400000-kickstarter-projects#kickstarter

Customized Feature Overview

General Information:
→ *duration, funding amount, category, location, description length*

Lexical Richness: the quality of vocabulary in a language sample
→ *lexical diversity, lexical entropy*

Lexical Readability: the ease with which a reader can understand a written text
→ *Flesch score, Gunning fog index, Flesch Kincaid index, Dale Chall readability score, text standard score*

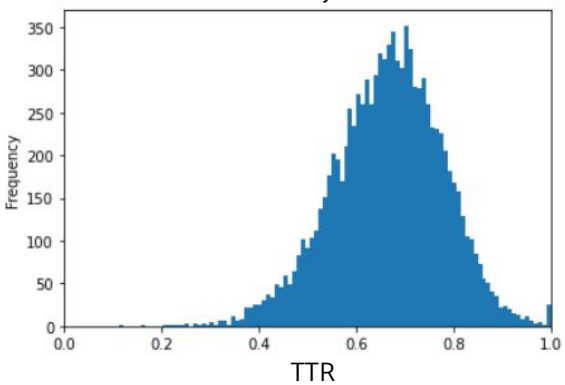
Reference: Zhou, Jamie, etc. (2015). Money talks: A predictive model on crowdfunding success using project description

Lexical Richness

Lexical diversity

Type Token Ratio (TTR) = $\dfrac{\text{Total number of UNIQUE tokens}}{\text{Total number of tokens}}$

Lexical diversity distribution

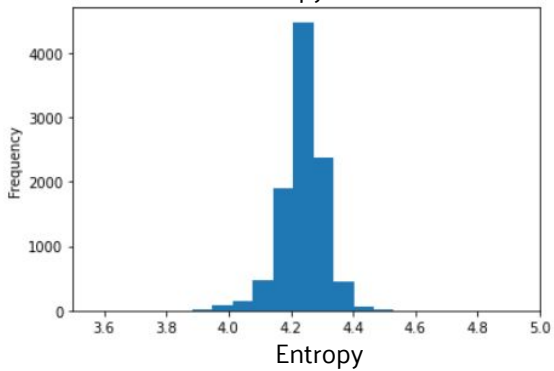


Lexical entropy

$$H = -100 \frac{\sum_i p_i \log p_i}{\log N}$$

where N = total number of tokens,
pi = probability of appearance for token i

Lexical entropy distribution



1. Thoiron, Philippe. “Diversity Index and Entropy as Measures of Lexical Richness.” Computers and the Humanities, vol. 20, no. 3, 1986, pp. 197–202
2. Dale, Moisl, and Somers (p.551). ”Handbook of Natural Language Processing” (2000)

Lexical Readability

Flesch reading ease score:
Higher scores indicate material that is easier to read

$$206.835 - 1.015\left(\frac{\text{total words}}{\text{total sentences}}\right) - 84.6\left(\frac{\text{total syllables}}{\text{total words}}\right)$$

Flesch Kincaid grade index:
Number of years of education required to understand

$$0.39\left(\frac{\text{total words}}{\text{total sentences}}\right) + 11.8\left(\frac{\text{total syllables}}{\text{total words}}\right) - 15.59$$

Gunning FOG index:
Number of years of formal education needed to understand

$$0.4\left[\left(\frac{\text{words}}{\text{sentences}}\right) + 100\left(\frac{\text{complex words}}{\text{words}}\right)\right]$$

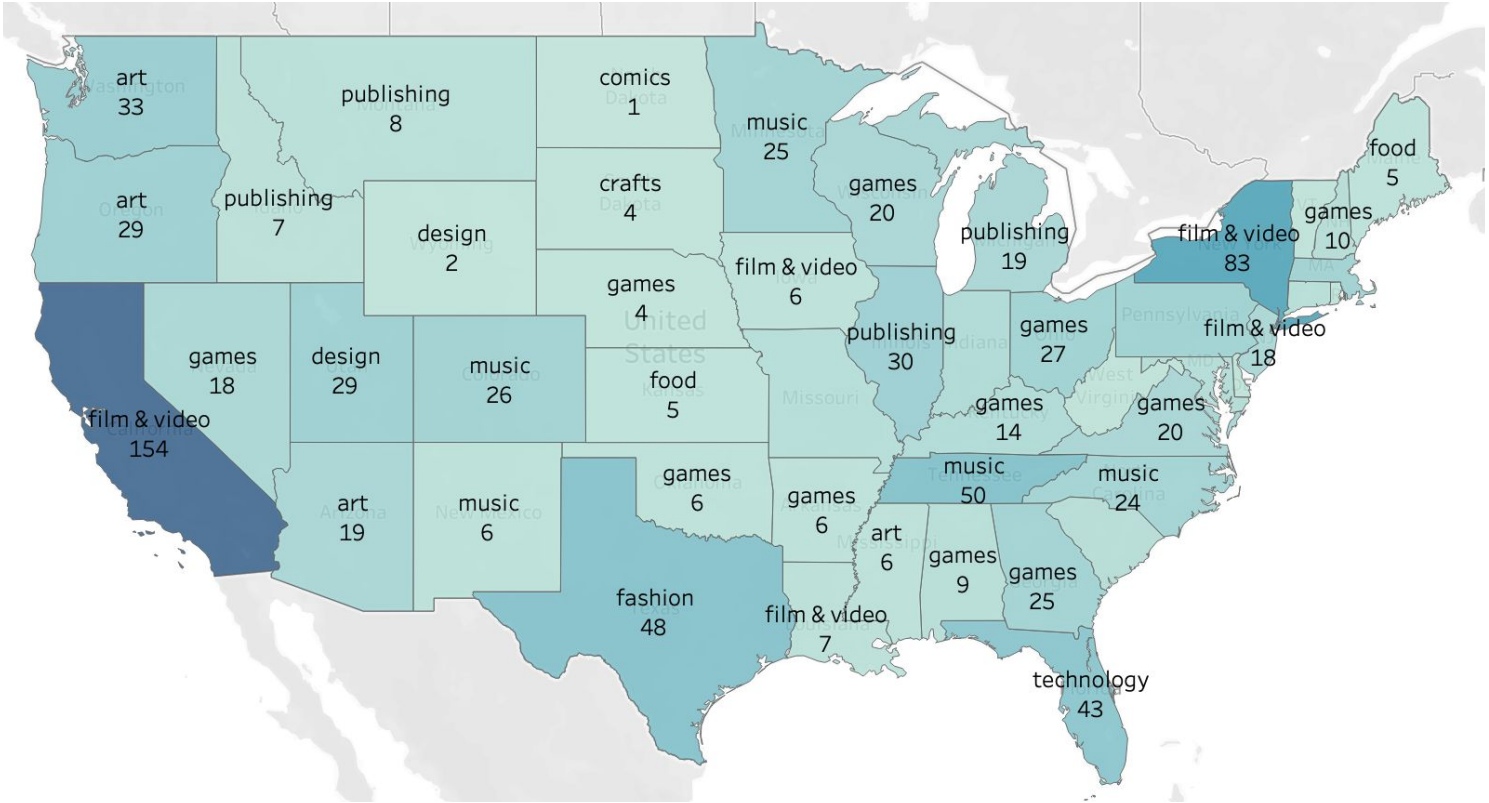
Dale Chall readability score:
Comprehension difficulty based on 3000 common words

$$0.1579\left(\frac{\text{difficult words}}{\text{words}} \times 100\right) + 0.0496\left(\frac{\text{words}}{\text{sentences}}\right)$$

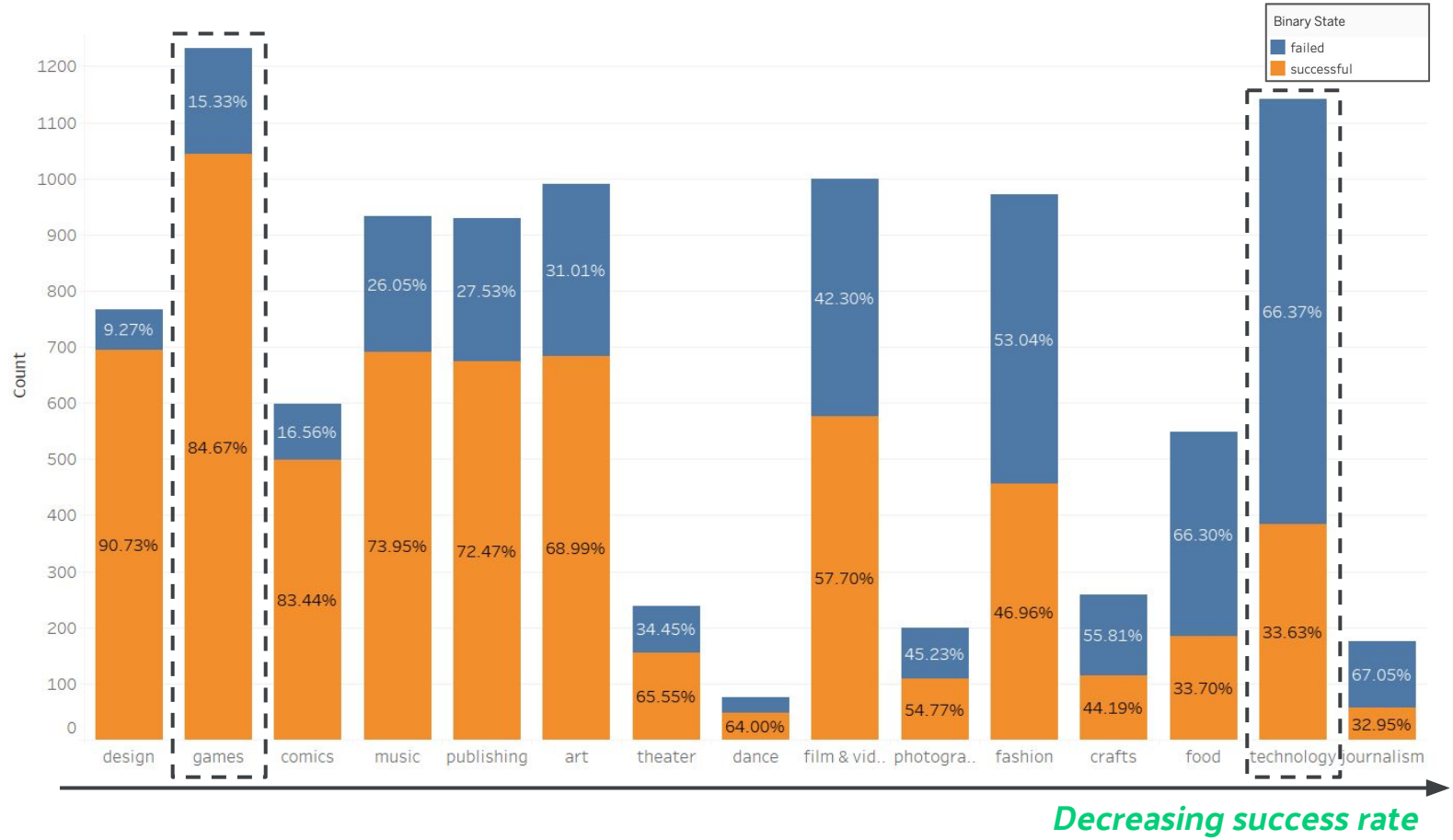
Text standard score:
Ensembled estimation of school grade level required to understand the text

References: 1. https://pypi.org/project/textstat/; 2. https://en.wikipedia.org/wiki/Flesch%E2%80%93Kincaid_readability_tests#Flesch_reading_ease;
3. https://en.wikipedia.org/wiki/Gunning_fog_index; 4. https://en.wikipedia.org/wiki/Dale%E2%80%93Chall_readability_formula

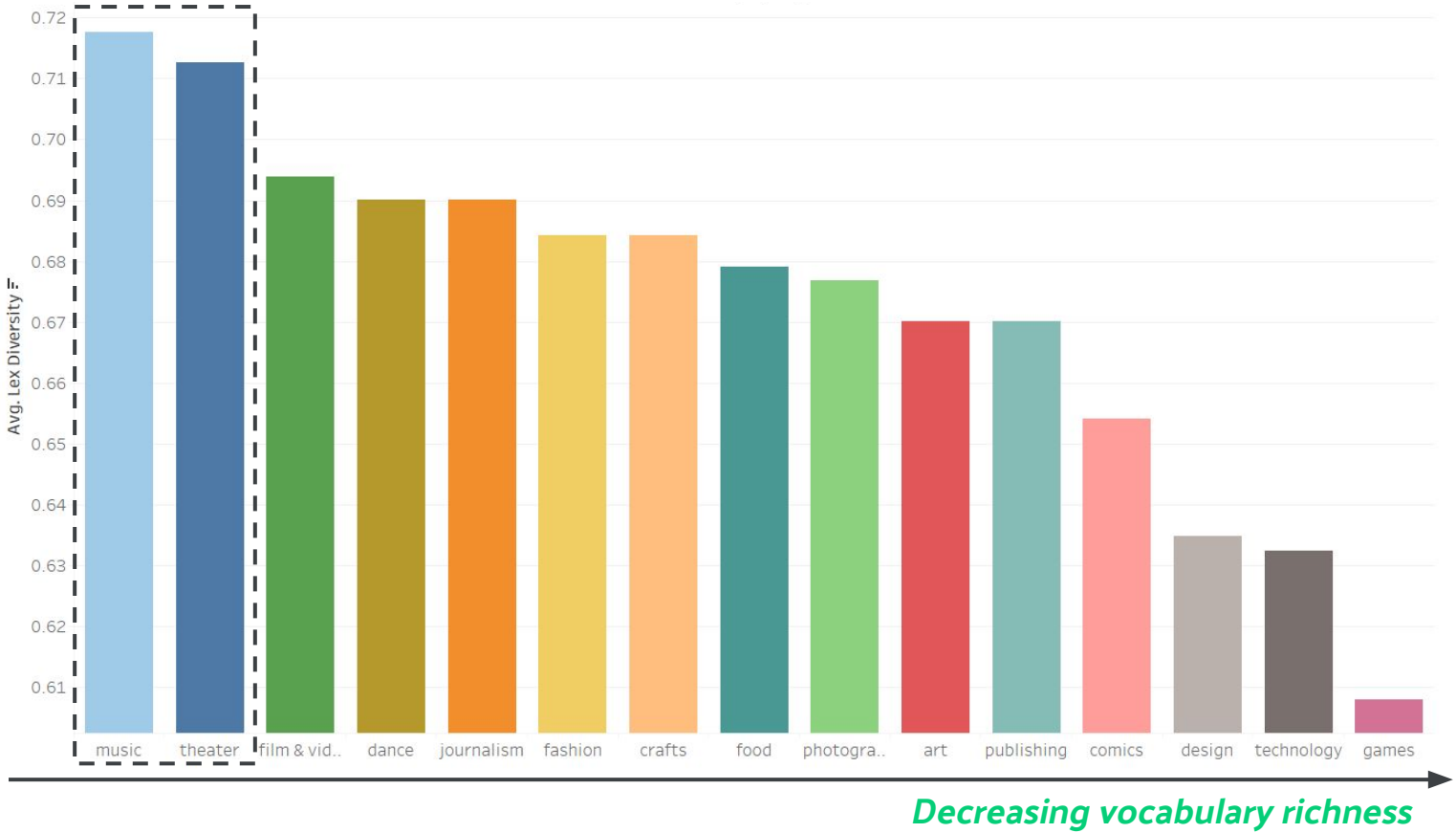
Most Kickstarter Categories by State



Project Number & Success Rate by Category

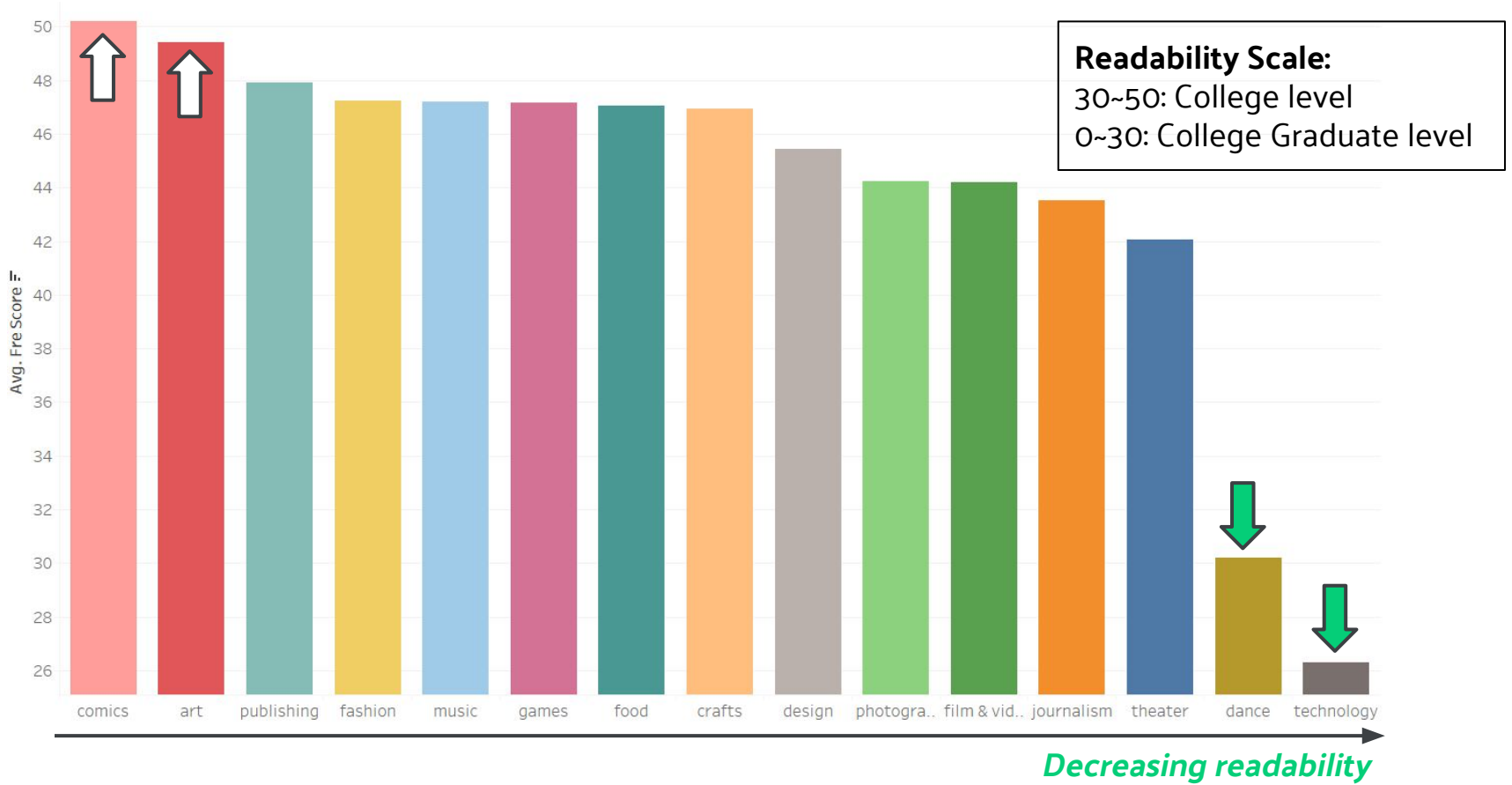


Lexical Diversity by Category



* Lexical diversity is measured by Type Token Ratio (TTR). Higher the lexical diversity score, richer the vocabulary

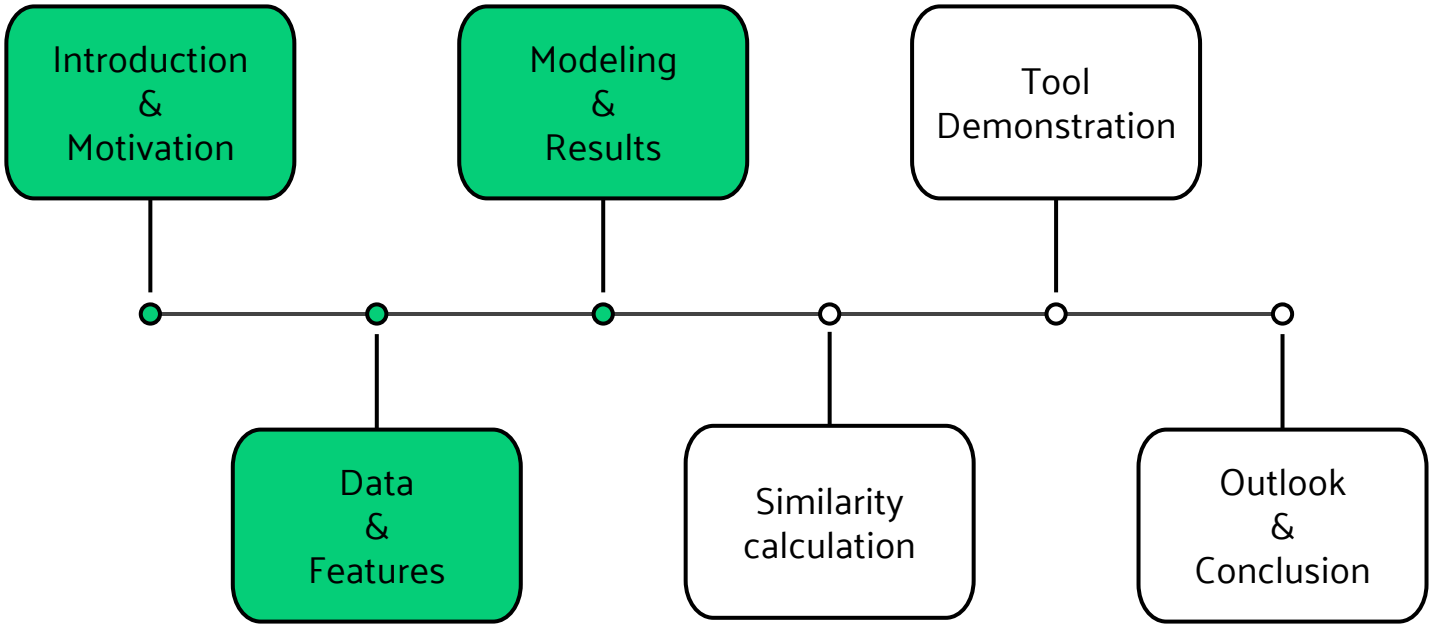
Lexical Readability by Category



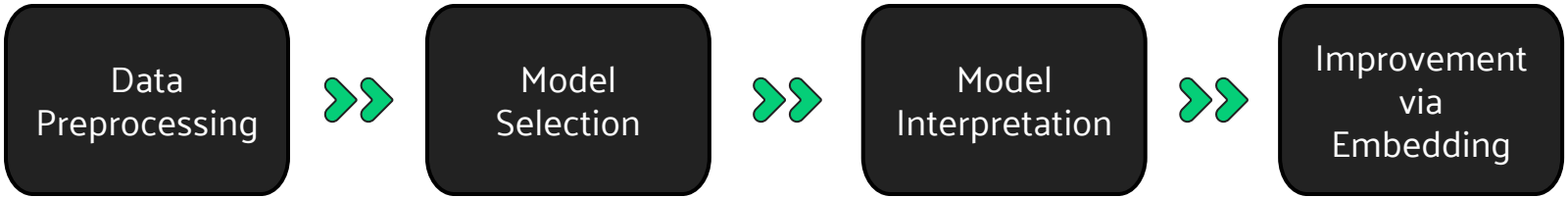
Readability Scale:
30~50: College level
0~30: College Graduate level

*Fre score = Flesch reading ease score. Higher the score, easier to read

Presentation Outline

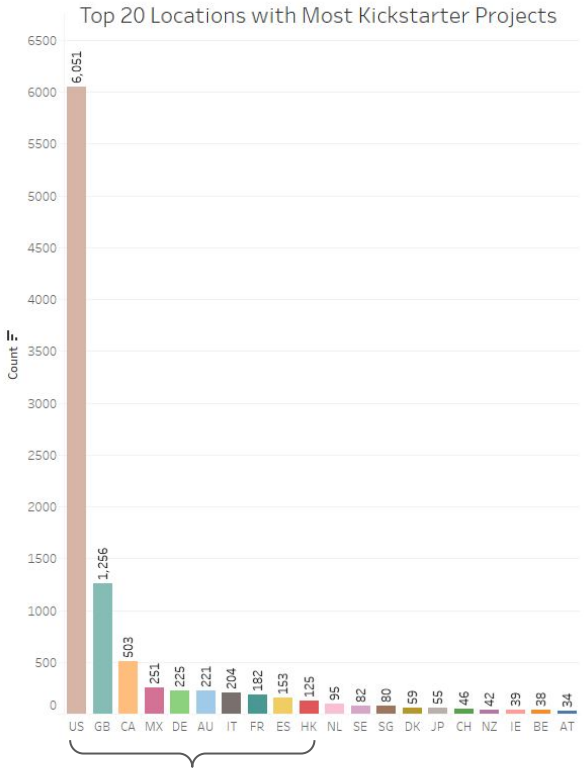


Modeling Process Overview

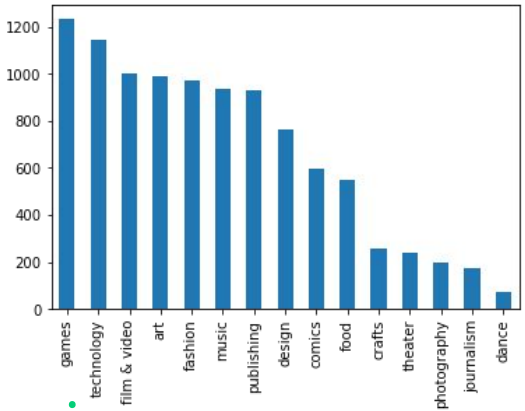


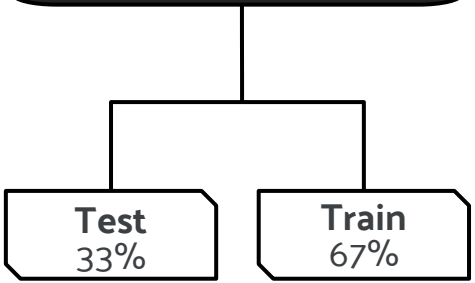
Data Preprocessing Steps

Location Transform » **Category Transform** » **Train/Test Split**



top_ten = 1; otherwise, *top_ten* = 0





Converted to (k-1) dummy variables

name	art	comics	crafts	design	fashion	film & video	food	games	journalism	music	photography	publishing	technology	theater
4tec: a New 3D Connect 4 Game	0	0	0	0	0	0	0	1	0	0	0	0	0	0
Komrad - Community | Care | Comradery	0	0	0	0	0	0	0	0	0	0	0	0	1	0
Jonesy on the Go Expansion Project	0	0	0	0	0	0	1	0	0	0	0	0	0	0
NOT WATTS- The restaurant that runs without electricity.	0	0	0	0	0	0	1	0	0	0	0	0	0	0
Healthy, fresh, delicious food delivered to your doorstep!	0	0	0	0	0	0	1	0	0	0	0	0	0	0

Data Preprocessing Results

name	result	days_to_deadline	goal_USD	top_ten	length	lex_diversity	lex_entropy	fre_score	fog_index	fkg_index	...	fashion	film & video	food	games	journalism	music	photography	publishing	technology	theater
4tec: a New 3D Connect 4 Game	1.0	30	5781.91765	0	2065	0.595870	4.273365	44.48	21.98	19.9	...	0	0	0	1	0	0	0	0	0	0
Komrad - Community | Care | Comradery	0.0	30	77932.17900	1	711	0.763441	4.247977	45.90	13.79	13.1	...	0	0	0	0	0	0	0	0	1	0
Jonesy on the Go Expansion Project	0.0	60	5000.00000	1	1342	0.711957	4.218493	42.14	16.54	14.6	...	0	0	1	0	0	0	0	0	0	0
NOT WATTS- The restaurant that runs without electricity.	0.0	30	9687.54885	1	949	0.765152	4.194116	34.12	18.52	17.6	...	0	0	1	0	0	0	0	0	0	0
Healthy, fresh, delicious food delivered to your doorstep!	0.0	30	3162.30532	1	1060	0.779310	4.238193	52.70	11.93	10.5	...	0	0	1	0	0	0	0	0	0	0

Train: 6739 records; Test: 3319 records

Label: 1 = successful; 0 = failed

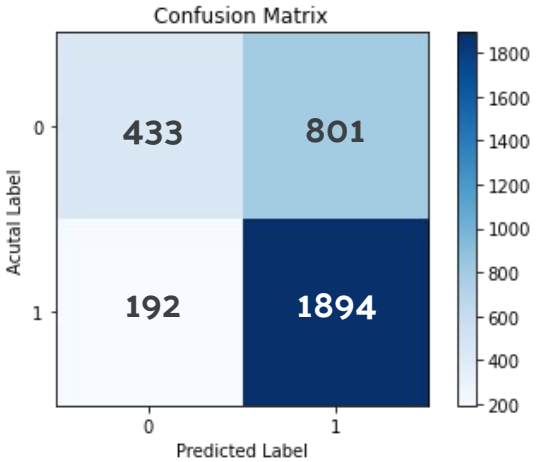
25 **Customized features**

Model Results

Choice of evaluation matrix:

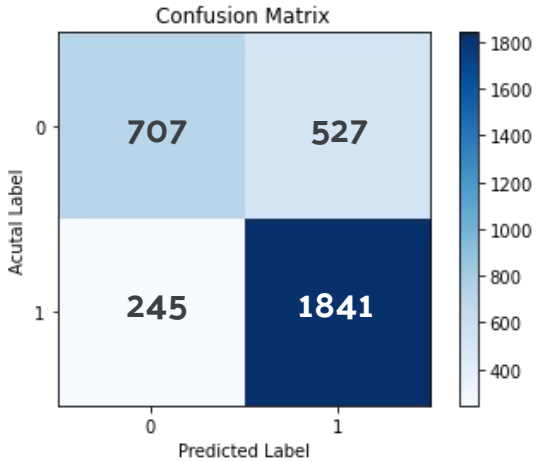
- **F1 score:** evaluate the overall model performance
- **Recall:** higher recall indicates lower chance of missed opportunity

Logistic Regression



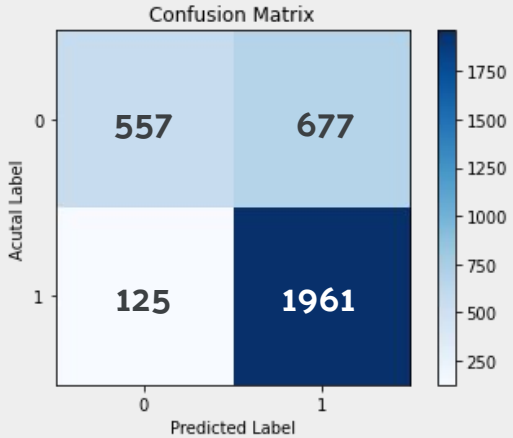
F1 score: 0.79
Recall score: 0.90

Random Forest



F1 score: 0.826
Recall score: 0.88

XGBoosting

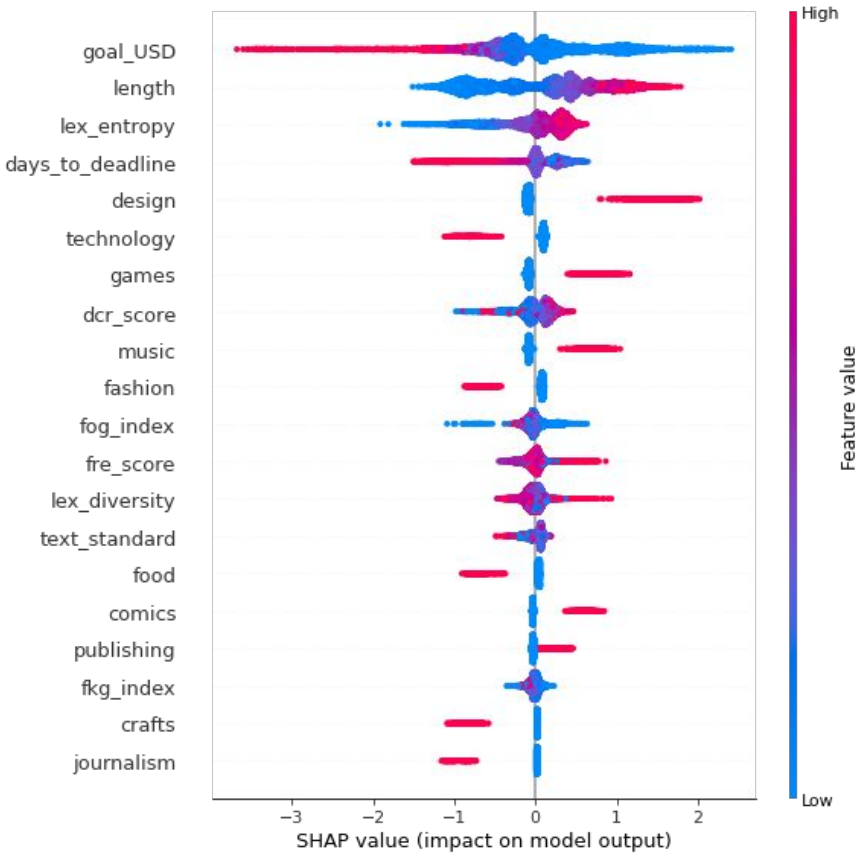


F1 score: 0.83
Recall score: 0.94

Better model performance

* Results from testing data; parameters are tuned via GridSearch; 1 = successful, 0 = failed

Model Interpretation - Feature Importance



General Practice for Successful Projects

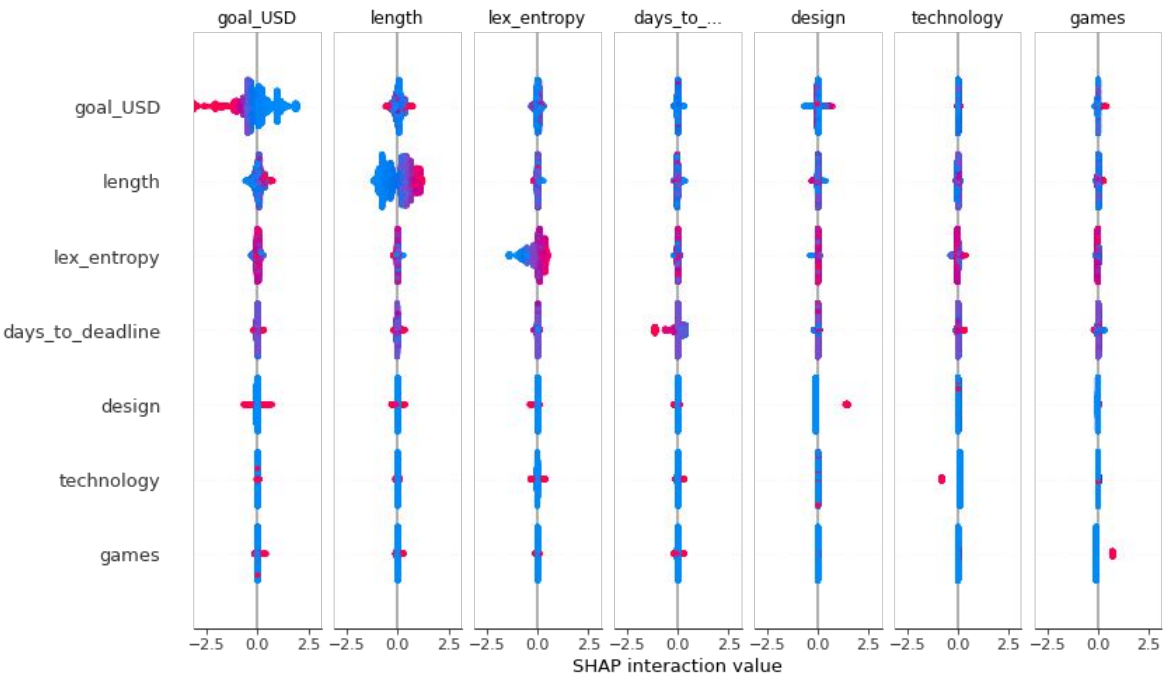
- **Lower** target funding amount
- **Shorter** funding collection time
- **Longer** story/product description
- **Richer** vocabulary in writing
- **Less** professional terminology/jargon

Category Specific Trends

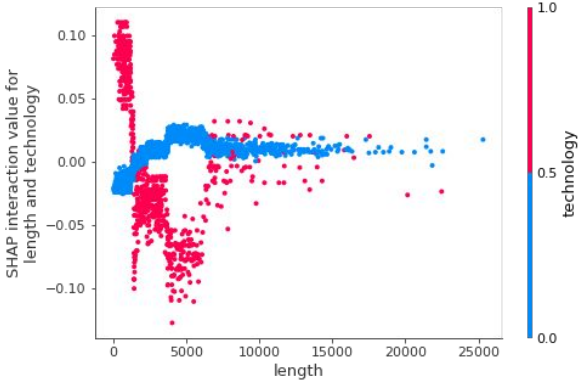
- **Technology, fashion, food, crafts,** and **journalism** projects are harder to get fully funded on Kickstarter
- **Design, games, music, comics,** and **publishing** projects are easier to get fully funded on Kickstarter

Reference: Lundberg, S.M., Erion, G., Chen, H. *et al.* From local explanations to global understanding with explainable AI for trees. *Nat Mach Intell* 2, 56–67 (2020). https://doi.org/10.1038/s42256-019-0138-9

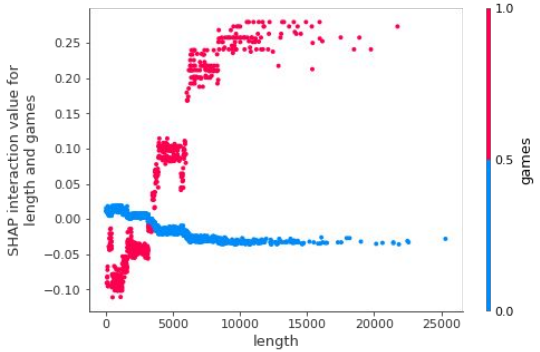
Model Interpretation - Feature Interactions



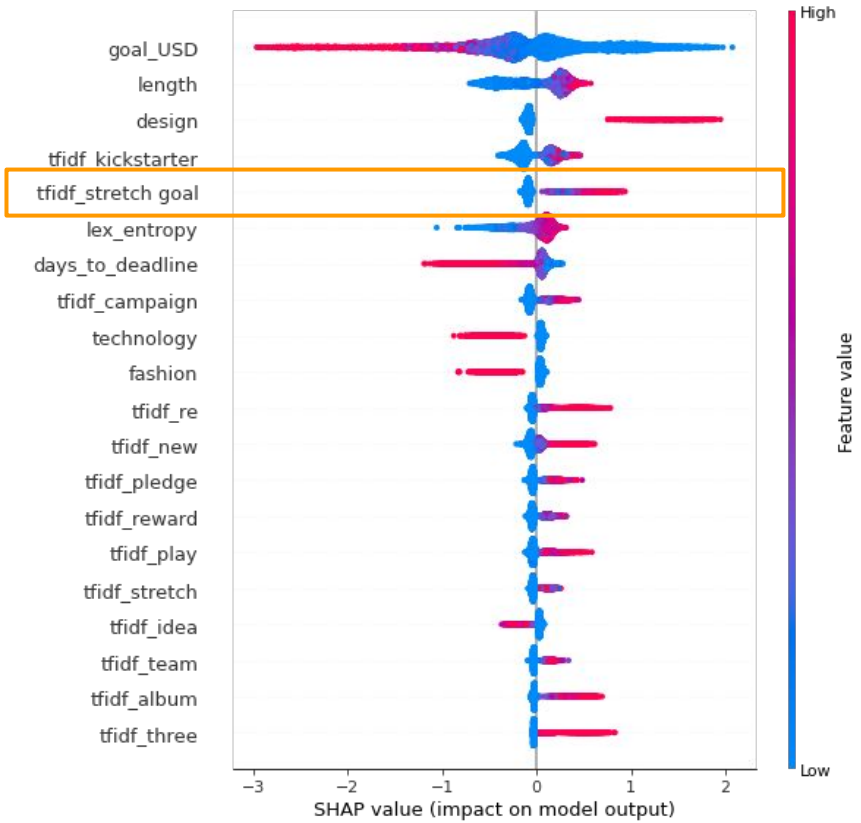
Technology: better to have shorter story



Games: better to have longer story



Important TF-IDF Terms For Prediction



Word cloud for successful projects

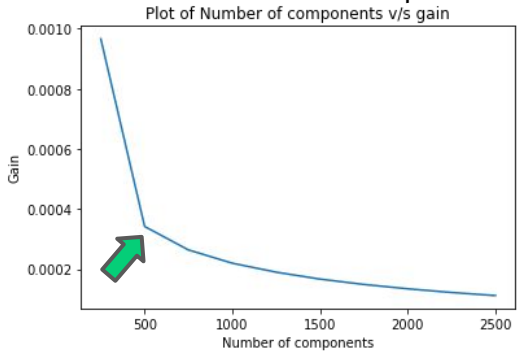
- Feature importance of the model with customized features and tf-idf features
- “Stretch goal” is an important term for project success
- Stretch goal refers to additional goals after the project is successfully funded
- Projects with stretch goals (i.e. future plans) are more likely to be funded.

* Results from XGBoosting model with 25 customized features and 40,000 tf-idf vector features

Model Results With Word Embedding

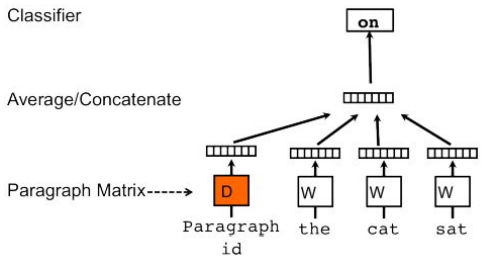
LSA

40K tfidf vectors → 500 components



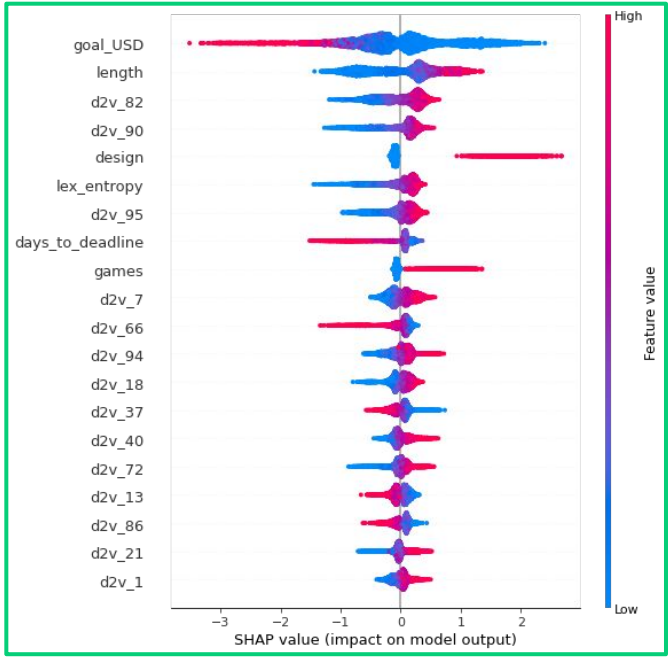
Doc2Vec

PV-DM Model



Source: https://medium.com/wisio/a-gentle-introduction-to-doc2vec-db3e8c0cce5e

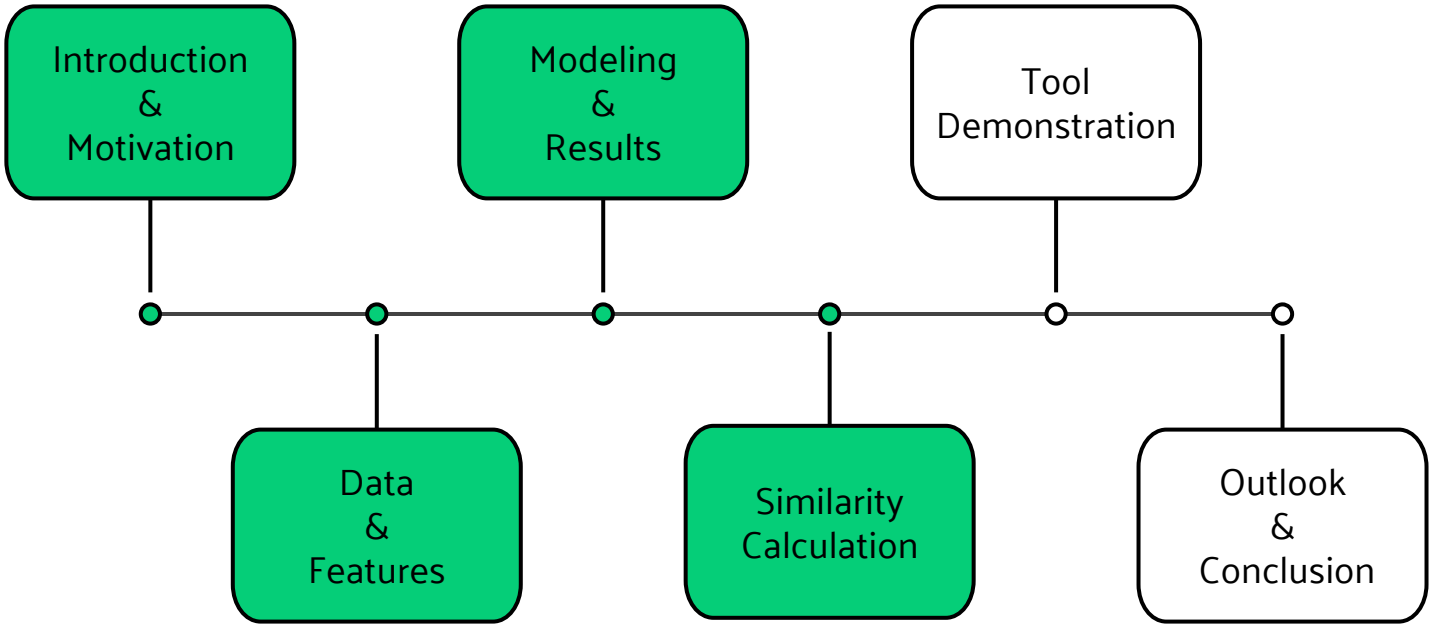
Cust Feat + Doc2Vec Model Results



Features	F1 score	Recall Score
Customized (control)	0.83	0.94
Customized + LSA	0.84	0.92
Customized + Doc2Vec	**0.85**	**0.92**
Doc2Vec (control)	0.81	0.90

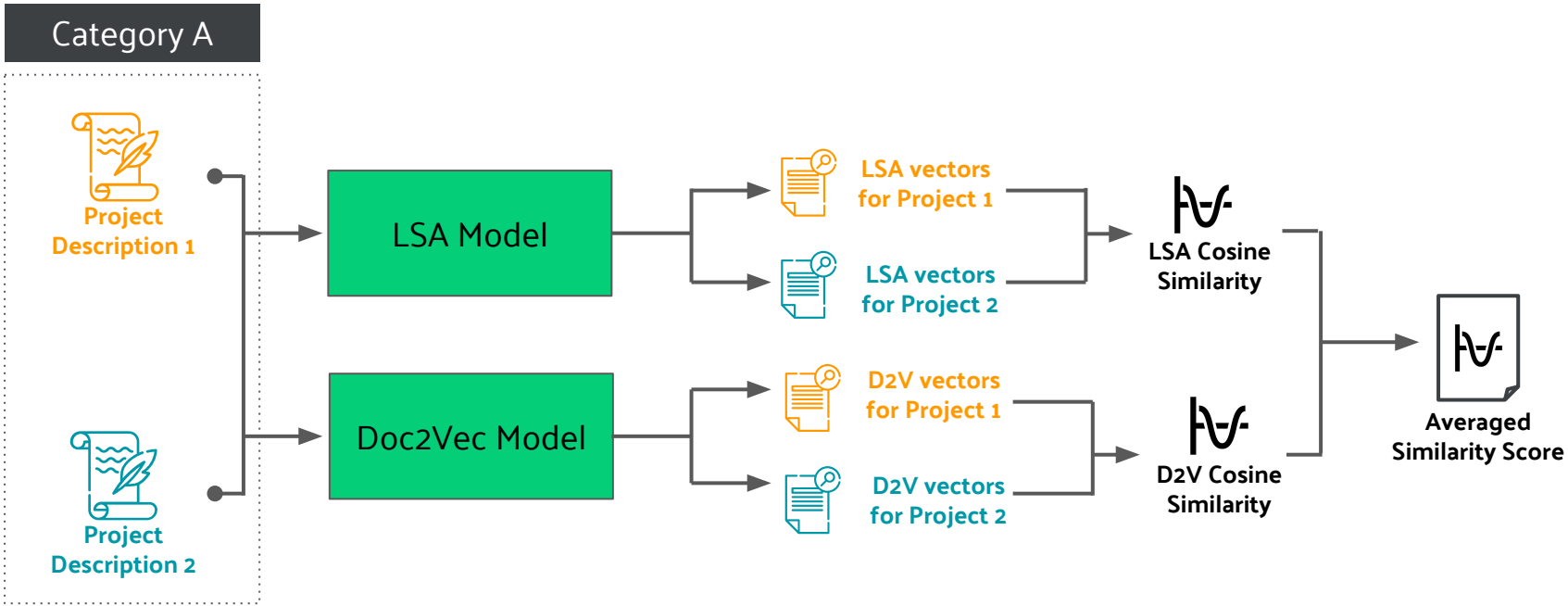
References: 1. A gentle introduction to Doc2Vec. https://medium.com/wisio/a-gentle-introduction-to-doc2vec-db3e8c0cce5e

Presentation Outline



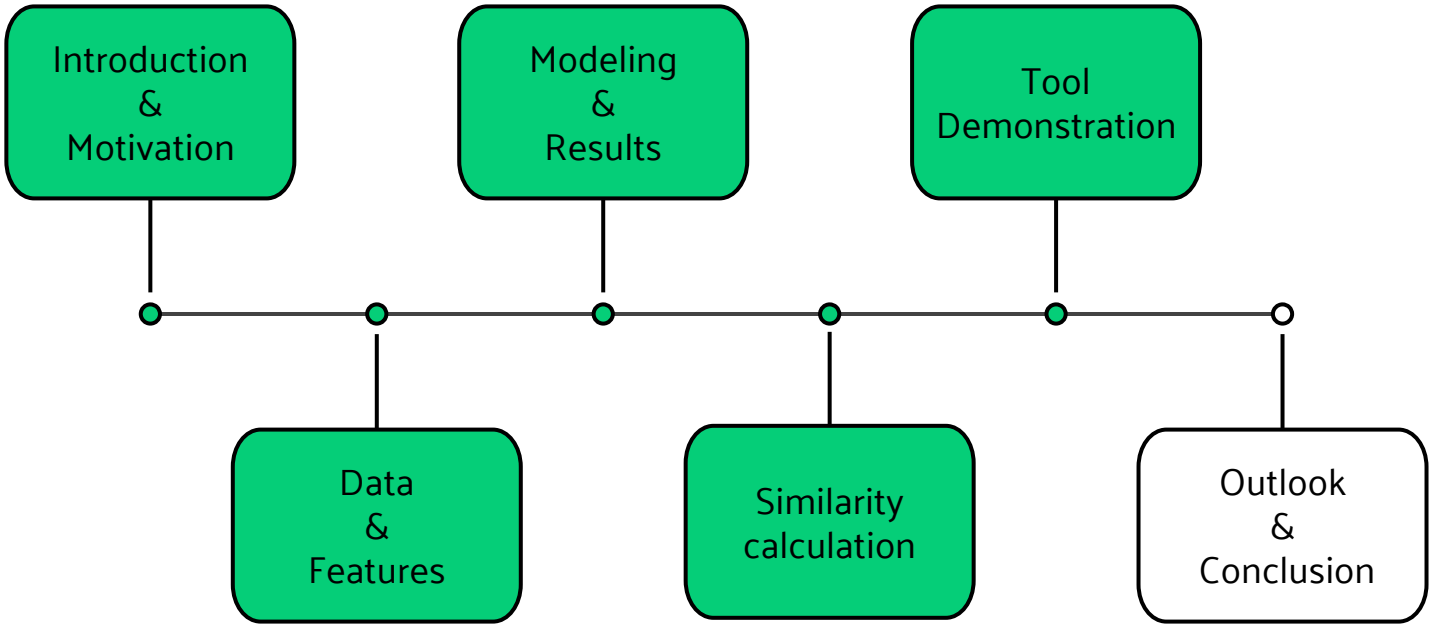
Similarity Score Calculation

Motivation: with user's input to identify similar Kickstarter projects from same category



References: Perfume recommendations using NLP. https://towardsdatascience.com/perfume-recommendations-using-natural-language-processing-ad3e6736074c

Presentation Outline



Design & Workflow

User inputs information

Description: e.g. 'I have a great idea...'

Category: games

Amount: e.g. '5000'

Duration: e.g. '21'

Submit

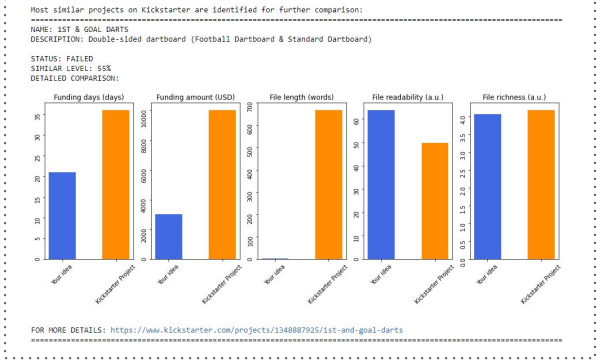
Modular & Automated

1. Input Preprocessing
2. Feature Creation
3. Make Prediction
4. Calculate Similarity

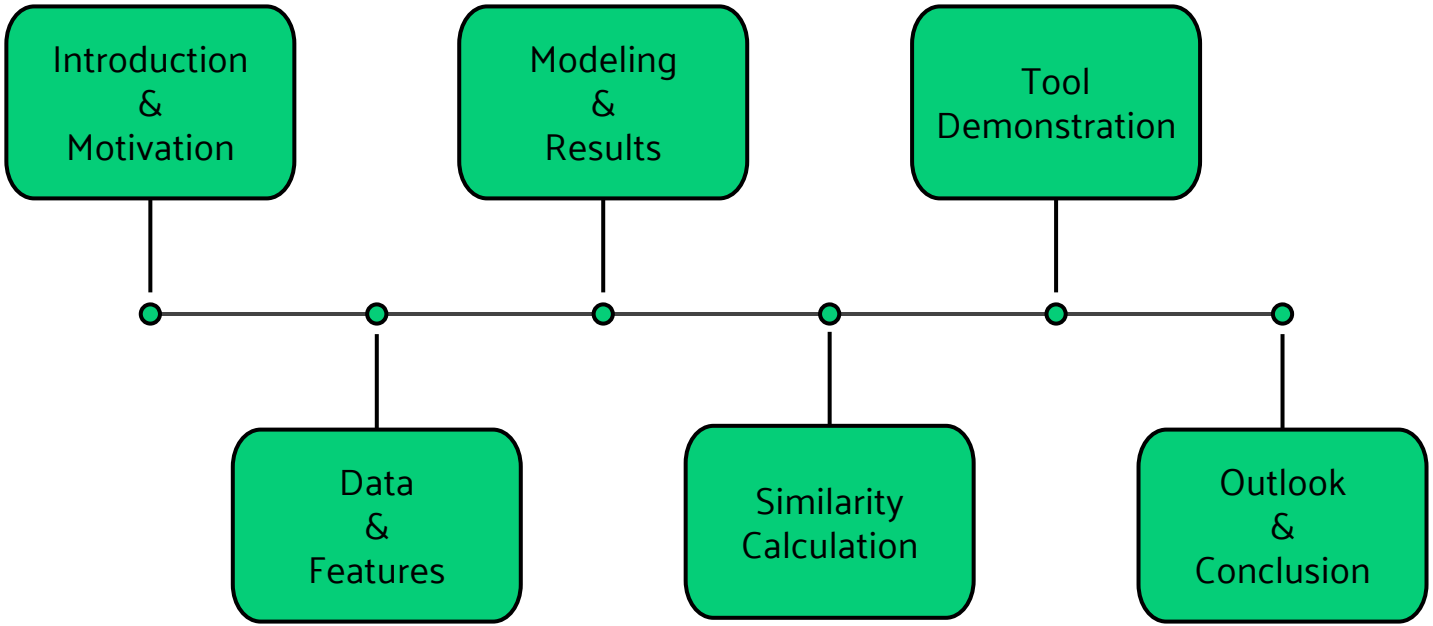
Output Prediction Results

```
==================================================
KICKSTARTER IDEA VALIDATOR RESULTS:
==================================================

Based on your input, the forecasted results are:
=> Predicted results: SUCCESSFUL
=> Successful probability: 54%
==================================================
```

Find and Compare with Similar projects



Presentation Outline



Summary & Future Work

What we did

- Explored the customized features from general information, lexical richness and lexical readability to build up the predictive model for Kickstarter projects
- Improved model performance via Doc2Vec word embedding method
- Used the ensemble similarity score to search for similar projects
- Integrated above-mentioned functions into a Kickstarter idea validation tool

What we found

- General practice for successful projects (**Lower** target funding amount, **Shorter** funding collection time, **Longer** story/product description, **Richer** vocabulary in writing, **Less** professional terminology)
- Category specific trends (**hardly funded**: Technology, fashion, food, crafts, and journalism; **easily funded: Design, games, music, comics, and publishing**)

How we can improve

- Increase dataset size via web scraping
- Use Kickstarter user-related information for modeling (e.g. previous project history, individual vs. company, etc), if accessible


THANK YOU


KICKSTARTER
work
game
time
book
one