CAPSTONE 2: PREDICTING KICKSTARTER CAMPAIGN OUTCOMES USING NLP

Bazeley,
Mikiko
Springboard – May 2019

TABLE OF CONTENTS

PROJECT OVERVIEW

DATA OVERVIEW

DATA ACQUISITION & PROCESSING

FEATURE ENGINEERING & SELECTION

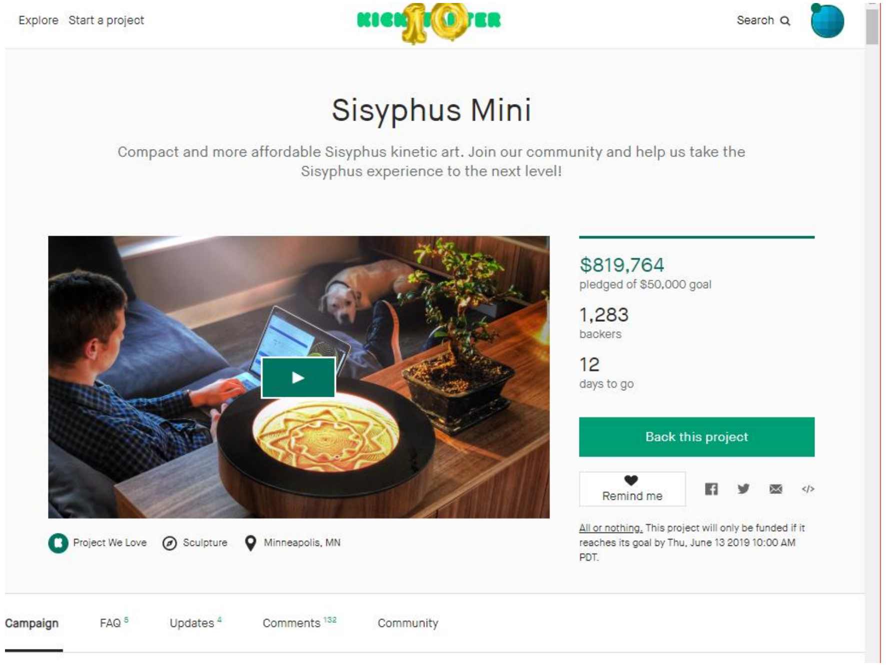
MODEL SELECTION & PERFORMANCE

TAKE-AWAYS

PROBLEM OVERVIEW

- Goal:
  - Predict campaign success on Kickstarter by leveraging NLP techniques.
- Potential Interested Parties:
  - Potential creators wanting to understand how well their campaign will perform
  - Competitive crowdfunding sites wanting to understand drivers of successful campaigns
- Data Source:
  - CSV files hosted on Web Robots, site which scrapes Kickstarter monthly

- Outcome:
  - Create predictive model utilizing numeric, text and categorical data in order to predict whether a campaign is successful within 15 days of being launched.



Sisyphus Mini

Compact and more affordable Sisyphus kinetic art. Join our community and help us take the Sisyphus experience to the next level!

$819,764 pledged of $50,000 goal

1,283 backers

12 days to go

Project We Love · Sculpture · Minneapolis, MN

All or nothing. This project will only be funded if it reaches its goal by Thu, June 13 2019 10:00 AM PDT.

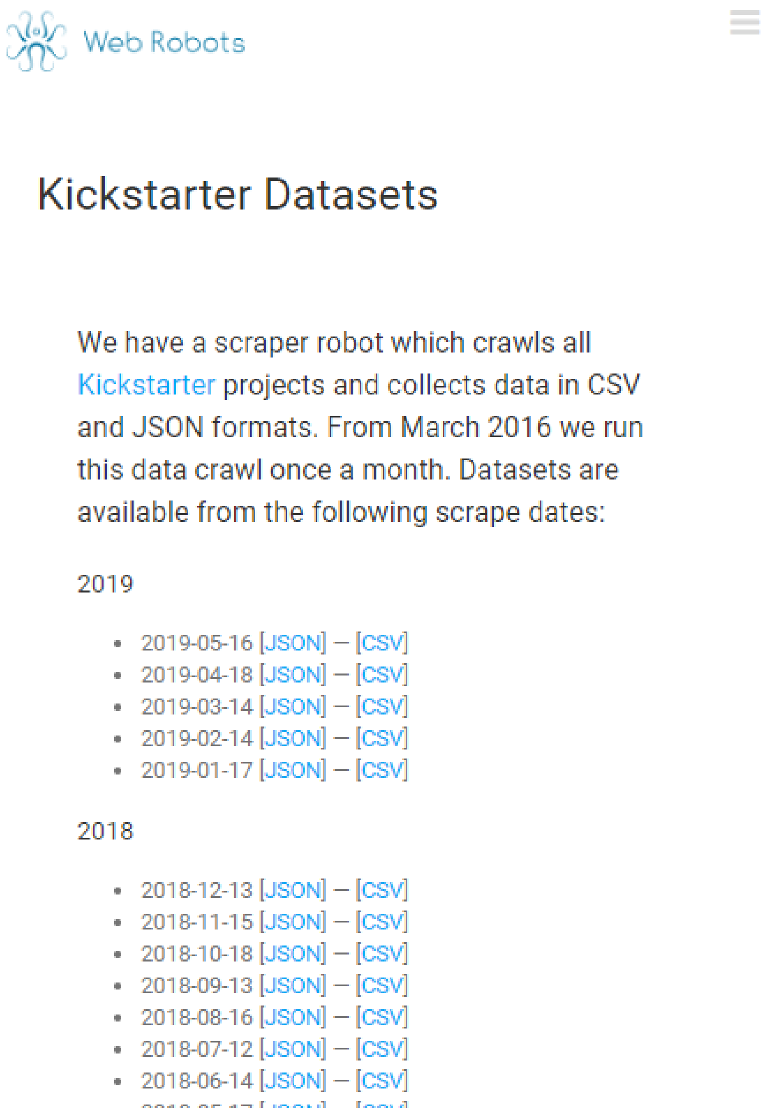
Example of a Kickstarter campaign – one of the most successful in Kickstarter history!

DATA OVERVIEW

- ❖Series of CSV files posted to site hosted by webrobots, consisting of monthly scrapes from April 22, 2019 to current month
- ❖Each scrape produced 25=50+ csv files, stored in folders corresponding to the date scraped.
- ❖Single row represents single campaign, with each scrape being a poir in time snapshot of the project
- ❖For this project, used data from Jan 28, 2016 to April 2019

❖Available Features:

- ❖'backers_count', 'blurb', 'category',
- ❖ 'converted_pledged_amount', 'country', 'created_at', 'creator',
- ❖ 'currency', 'currency_symbol', 'currency_trailing_code',
- ❖ 'current_currency', 'deadline', 'dirname', 'disable_communication',
- ❖ 'friends', 'fx_rate', 'goal', 'id', 'is_backing', 'is_starrable',
- ❖ 'is_starred', 'last_update_published_at', 'launched_at', 'location',
- ❖ 'name', 'permissions', 'photo', 'pledged', 'profile', 'slug',
- ❖ 'source_url', 'spotlight', 'staff_pick', 'state', 'state_changed_at',
- ❖ 'static_usd_rate', 'unread_messages_count', 'unseen_activity_count',
- ❖ 'urls', 'usd_pledged', 'usd_type'



Kickstarter Datasets

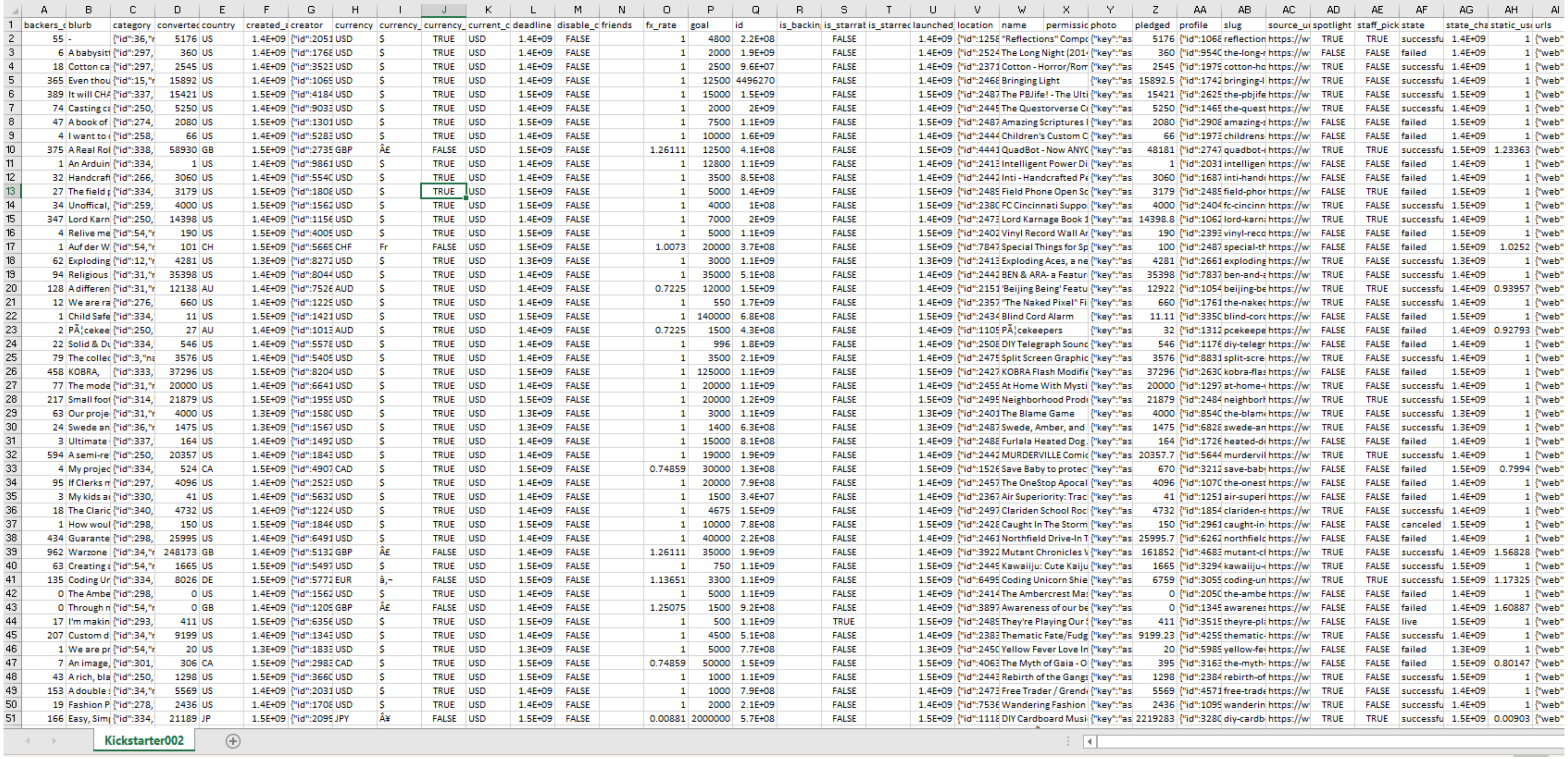
We have a scraper robot which crawls all Kickstarter projects and collects data in CSV and JSON formats. From March 2016 we run this data crawl once a month. Datasets are available from the following scrape dates:

2019

- 2019-05-16 [JSON] – [CSV]
- 2019-04-18 [JSON] – [CSV]
- 2019-03-14 [JSON] – [CSV]
- 2019-02-14 [JSON] – [CSV]
- 2019-01-17 [JSON] – [CSV]

2018

- 2018-12-13 [JSON] – [CSV]
- 2018-11-15 [JSON] – [CSV]
- 2018-10-18 [JSON] – [CSV]
- 2018-09-13 [JSON] – [CSV]
- 2018-08-16 [JSON] – [CSV]
- 2018-07-12 [JSON] – [CSV]
- 2018-06-14 [JSON] – [CSV]

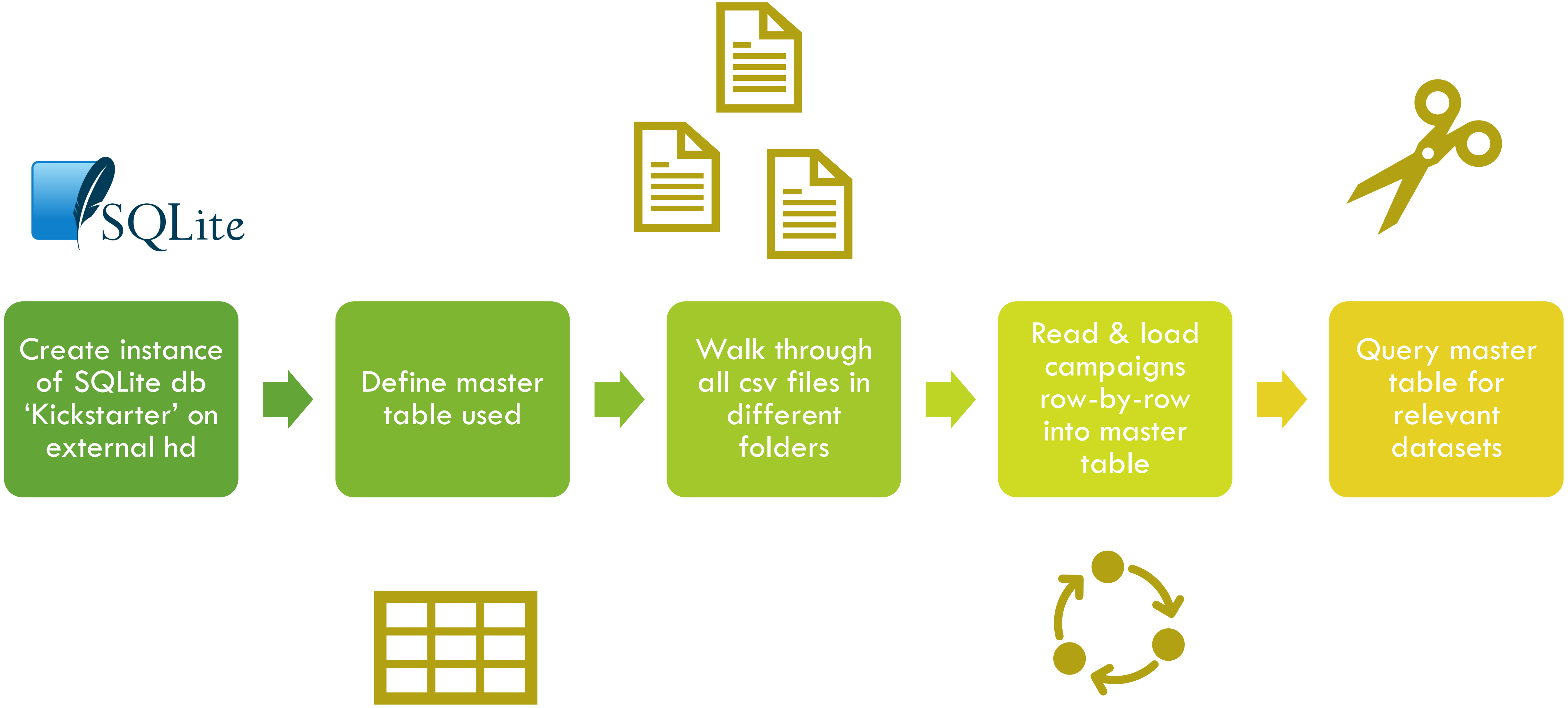


	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC	AD	AE	AF	AG	AH	AI
1	backers_c	blurb	category	convertec	country	created_a	creator	currency	currency_	currency_	current_c	deadline	disable_c	friends	fx_rate	goal	id	is_backin	is_starrat	is_starrec	launched	location	name	permissic	photo	pledged	profile	slug	source_ur	spotlight	staff_pick	state	state_cha	static_usc	urls
2	55	-	{"id":36,"r	5176	US	1.4E+09	{"id":2051	USD	$	TRUE	USD	1.4E+09	FALSE		1	4800	2.2E+08		FALSE		1.4E+09	{"id":1258	"Reflections" Compo		{"key":"as	5176	{"id":1068	reflection	https://w	TRUE	TRUE	successfu	1.4E+09	1	{"web"
3	6	A babysitt	{"id":297,	360	US	1.4E+09	{"id":1768	USD	$	TRUE	USD	1.4E+09	FALSE		1	2000	1.9E+09		FALSE		1.4E+09	{"id":2524	The Long Night (2014		{"key":"as	360	{"id":9540	the-long-r	https://w	FALSE	FALSE	failed	1.4E+09	1	{"web"
4	18	Cotton ca	{"id":297,	2545	US	1.4E+09	{"id":3523	USD	$	TRUE	USD	1.4E+09	FALSE		1	2500	9.6E+07		FALSE		1.4E+09	{"id":2371	Cotton - Horror/Rom		{"key":"as	2545	{"id":1979	cotton-ho	https://w	TRUE	FALSE	successfu	1.4E+09	1	{"web"
5	365	Even thou	{"id":15,"r	15892	US	1.4E+09	{"id":1069	USD	$	TRUE	USD	1.4E+09	FALSE		1	12500	4496270		FALSE		1.4E+09	{"id":2468	Bringing Light		{"key":"as	15892.5	{"id":1742	bringing-l	https://w	TRUE	FALSE	successfu	1.4E+09	1	{"web"
6	389	It will CHA	{"id":337,	15421	US	1.5E+09	{"id":4184	USD	$	TRUE	USD	1.5E+09	FALSE		1	15000	1.5E+09		FALSE		1.5E+09	{"id":2487	The PBJife! - The Ulti		{"key":"as	15421	{"id":2625	the-pbjife	https://w	TRUE	FALSE	successfu	1.5E+09	1	{"web"
7	74	Casting ca	{"id":250,	5250	US	1.4E+09	{"id":9033	USD	$	TRUE	USD	1.4E+09	FALSE		1	2000	2E+09		FALSE		1.4E+09	{"id":2445	The Questorverse C		{"key":"as	5250	{"id":1465	the-quest	https://w	TRUE	FALSE	successfu	1.4E+09	1	{"web"
8	47	A book of	{"id":274,	2080	US	1.5E+09	{"id":1301	USD	$	TRUE	USD	1.5E+09	FALSE		1	7500	1.1E+09		FALSE		1.5E+09	{"id":2487	Amazing Scriptures I		{"key":"as	2080	{"id":2908	amazing-s	https://w	FALSE	FALSE	failed	1.5E+09	1	{"web"
9	4	I want to	{"id":258,	66	US	1.4E+09	{"id":5283	USD	$	TRUE	USD	1.4E+09	FALSE		1	10000	1.6E+09		FALSE		1.4E+09	{"id":2444	Children's Custom C		{"key":"as	66	{"id":1973	childrens-	https://w	FALSE	FALSE	failed	1.4E+09	1	{"web"
10	375	A Real Rol	{"id":338,	58930	GB	1.5E+09	{"id":2735	GBP	Â£	FALSE	USD	1.5E+09	FALSE		1.26111	12500	4.1E+08		FALSE		1.5E+09	{"id":4441	QuadBot - Now ANYC		{"key":"as	48181	{"id":2747	quadbot-	https://w	TRUE	TRUE	successfu	1.5E+09	1.23363	{"web"
11	1	An Arduin	{"id":334,	1	US	1.4E+09	{"id":9861	USD	$	TRUE	USD	1.4E+09	FALSE		1	12800	1.1E+09		FALSE		1.4E+09	{"id":2413	Intelligent Power Di		{"key":"as	1	{"id":2031	intelligen	https://w	FALSE	FALSE	failed	1.4E+09	1	{"web"
12	32	Handcraft	{"id":266,	3060	US	1.4E+09	{"id":5540	USD	$	TRUE	USD	1.4E+09	FALSE		1	3500	8.5E+08		FALSE		1.4E+09	{"id":2442	Inti - Handcrafted Pe		{"key":"as	3060	{"id":1687	inti-hand	https://w	FALSE	FALSE	failed	1.4E+09	1	{"web"
13	27	The field p	{"id":334,	3179	US	1.5E+09	{"id":1808	USD	$	TRUE	USD	1.5E+09	FALSE		1	5000	1.4E+09		FALSE		1.5E+09	{"id":2489	Field Phone Open Sc		{"key":"as	3179	{"id":2485	field-phor	https://w	FALSE	TRUE	failed	1.5E+09	1	{"web"
14	34	Unoffical,	{"id":259,	4000	US	1.5E+09	{"id":1562	USD	$	TRUE	USD	1.5E+09	FALSE		1	4000	1E+08		FALSE		1.5E+09	{"id":2380	FC Cincinnati Suppo		{"key":"as	4000	{"id":2404	fc-cincinn	https://w	TRUE	FALSE	successfu	1.5E+09	1	{"web"
15	347	Lord Karn	{"id":250,	14398	US	1.4E+09	{"id":1156	USD	$	TRUE	USD	1.4E+09	FALSE		1	7000	2E+09		FALSE		1.4E+09	{"id":2473	Lord Karnage Book 1		{"key":"as	14398.8	{"id":1062	lord-karn	https://w	TRUE	TRUE	successfu	1.4E+09	1	{"web"
16	4	Relive me	{"id":54,"r	190	US	1.5E+09	{"id":4005	USD	$	TRUE	USD	1.5E+09	FALSE		1	5000	1.1E+09		FALSE		1.5E+09	{"id":2402	Vinyl Record Wall Ar		{"key":"as	190	{"id":2393	vinyl-reco	https://w	FALSE	FALSE	failed	1.5E+09	1	{"web"
17	1	Auf der W	{"id":54,"r	101	CH	1.5E+09	{"id":5669	CHF	Fr	FALSE	USD	1.5E+09	FALSE		1.0073	20000	3.7E+08		FALSE		1.5E+09	{"id":7847	Special Things for Sp		{"key":"as	100	{"id":2487	special-th	https://w	FALSE	FALSE	failed	1.5E+09	1.0252	{"web"
18	62	Exploding	{"id":12,"r	4281	US	1.3E+09	{"id":8272	USD	$	TRUE	USD	1.3E+09	FALSE		1	3000	1.1E+09		FALSE		1.3E+09	{"id":2413	Exploding Aces, a ne		{"key":"as	4281	{"id":2661	exploding	https://w	TRUE	FALSE	successfu	1.3E+09	1	{"web"
19	94	Religious	{"id":31,"r	35398	US	1.4E+09	{"id":8044	USD	$	TRUE	USD	1.4E+09	FALSE		1	35000	5.1E+08		FALSE		1.4E+09	{"id":2442	BEN & ARA- a Featur		{"key":"as	35398	{"id":7837	ben-and-a	https://w	TRUE	FALSE	successfu	1.4E+09	1	{"web"
20	128	A differen	{"id":31,"r	12138	AU	1.4E+09	{"id":7526	AUD	$	TRUE	USD	1.4E+09	FALSE		0.7225	12000	1.5E+09		FALSE		1.4E+09	{"id":2151	'Beijing Being' Featu		{"key":"as	12922	{"id":1054	beijing-be	https://w	TRUE	TRUE	successfu	1.4E+09	0.93957	{"web"
21	12	We are ra	{"id":276,	660	US	1.4E+09	{"id":1229	USD	$	TRUE	USD	1.4E+09	FALSE		1	550	1.7E+09		FALSE		1.4E+09	{"id":2357	"The Naked Pixel" Fi		{"key":"as	660	{"id":1761	the-naked	https://w	TRUE	FALSE	successfu	1.4E+09	1	{"web"
22	1	Child Safe	{"id":334,	11	US	1.5E+09	{"id":1421	USD	$	TRUE	USD	1.5E+09	FALSE		1	140000	6.8E+08		FALSE		1.5E+09	{"id":2434	Blind Cord Alarm		{"key":"as	11.11	{"id":3350	blind-corc	https://w	FALSE	FALSE	failed	1.5E+09	1	{"web"
23	2	PÃ¦cekee	{"id":250,	27	AU	1.4E+09	{"id":1013	AUD	$	TRUE	USD	1.4E+09	FALSE		0.7225	1500	4.3E+08		FALSE		1.4E+09	{"id":1105	PÃ¦cekeepers		{"key":"as	32	{"id":1312	pcekeepe	https://w	FALSE	FALSE	failed	1.4E+09	0.92793	{"web"
24	22	Solid & Du	{"id":334,	546	US	1.4E+09	{"id":5578	USD	$	TRUE	USD	1.4E+09	FALSE		1	996	1.8E+09		FALSE		1.4E+09	{"id":2508	DIY Telegraph Sounc		{"key":"as	546	{"id":1176	diy-telegr	https://w	FALSE	FALSE	failed	1.4E+09	1	{"web"
25	79	The collec	{"id":3,"na	3576	US	1.4E+09	{"id":5405	USD	$	TRUE	USD	1.4E+09	FALSE		1	3500	2.1E+09		FALSE		1.4E+09	{"id":2475	Split Screen Graphic		{"key":"as	3576	{"id":8831	split-scre	https://w	TRUE	FALSE	successfu	1.4E+09	1	{"web"
26	458	KOBRA,	{"id":333,	37296	US	1.5E+09	{"id":8204	USD	$	TRUE	USD	1.5E+09	FALSE		1	125000	1.1E+09		FALSE		1.5E+09	{"id":2427	KOBRA Flash Modifie		{"key":"as	37296	{"id":2630	kobra-flas	https://w	FALSE	FALSE	failed	1.5E+09	1	{"web"
27	77	The mode	{"id":31,"r	20000	US	1.4E+09	{"id":6641	USD	$	TRUE	USD	1.4E+09	FALSE		1	20000	1.1E+09		FALSE		1.4E+09	{"id":2459	At Home With Mysti		{"key":"as	20000	{"id":1297	at-home-	https://w	TRUE	FALSE	successfu	1.4E+09	1	{"web"
28	217	Small foot	{"id":314,	21879	US	1.5E+09	{"id":1959	USD	$	TRUE	USD	1.5E+09	FALSE		1	20000	1.2E+09		FALSE		1.5E+09	{"id":2495	Neighborhood Prod		{"key":"as	21879	{"id":2484	neighborh	https://w	TRUE	TRUE	successfu	1.5E+09	1	{"web"
29	63	Our proje	{"id":31,"r	4000	US	1.3E+09	{"id":1580	USD	$	TRUE	USD	1.3E+09	FALSE		1	3000	1.1E+09		FALSE		1.3E+09	{"id":2401	The Blame Game		{"key":"as	4000	{"id":8540	the-blam	https://w	TRUE	FALSE	successfu	1.3E+09	1	{"web"
30	24	Swede an	{"id":36,"r	1475	US	1.3E+09	{"id":1567	USD	$	TRUE	USD	1.3E+09	FALSE		1	1400	6.3E+08		FALSE		1.3E+09	{"id":2487	Swede, Amber, and		{"key":"as	1475	{"id":6828	swede-an	https://w	TRUE	FALSE	successfu	1.3E+09	1	{"web"
31	3	Ultimate	{"id":337,	164	US	1.4E+09	{"id":1492	USD	$	TRUE	USD	1.4E+09	FALSE		1	15000	8.1E+08		FALSE		1.4E+09	{"id":2488	Furlala Heated Dog.		{"key":"as	164	{"id":1726	heated-d	https://w	FALSE	FALSE	failed	1.4E+09	1	{"web"
32	594	A semi-re	{"id":250,	20357	US	1.4E+09	{"id":1843	USD	$	TRUE	USD	1.4E+09	FALSE		1	19000	1.9E+09		FALSE		1.4E+09	{"id":2442	MURDERVILLE Comic		{"key":"as	20357.7	{"id":5644	murdervil	https://w	TRUE	TRUE	successfu	1.4E+09	1	{"web"
33	4	My projec	{"id":334,	524	CA	1.5E+09	{"id":4907	CAD	$	TRUE	USD	1.5E+09	FALSE		0.74859	30000	1.3E+08		FALSE		1.5E+09	{"id":1526	Save Baby to protec		{"key":"as	670	{"id":3212	save-baby	https://w	FALSE	FALSE	failed	1.5E+09	0.7994	{"web"
34	95	If Clerks m	{"id":297,	4096	US	1.4E+09	{"id":2523	USD	$	TRUE	USD	1.4E+09	FALSE		1	20000	7.9E+08		FALSE		1.4E+09	{"id":2457	The OneStop Apocal		{"key":"as	4096	{"id":1070	the-onest	https://w	FALSE	FALSE	failed	1.4E+09	1	{"web"
35	3	My kids ar	{"id":330,	41	US	1.4E+09	{"id":5632	USD	$	TRUE	USD	1.4E+09	FALSE		1	1500	3.4E+07		FALSE		1.4E+09	{"id":2367	Air Superiority: Trac		{"key":"as	41	{"id":1251	air-superi	https://w	FALSE	FALSE	failed	1.4E+09	1	{"web"
36	18	The Claric	{"id":340,	4732	US	1.4E+09	{"id":1224	USD	$	TRUE	USD	1.4E+09	FALSE		1	4675	1.5E+09		FALSE		1.4E+09	{"id":2497	Clariden School Roc		{"key":"as	4732	{"id":1854	clariden-s	https://w	TRUE	FALSE	successfu	1.4E+09	1	{"web"
37	1	How woul	{"id":298,	150	US	1.5E+09	{"id":1846	USD	$	TRUE	USD	1.5E+09	FALSE		1	10000	7.8E+08		FALSE		1.5E+09	{"id":2428	Caught In The Storm		{"key":"as	150	{"id":2961	caught-in-	https://w	FALSE	FALSE	canceled	1.5E+09	1	{"web"
38	434	Guarante	{"id":298,	25995	US	1.4E+09	{"id":6491	USD	$	TRUE	USD	1.4E+09	FALSE		1	40000	2.2E+08		FALSE		1.4E+09	{"id":2461	Northfield Drive-In T		{"key":"as	25995.7	{"id":6262	northfielc	https://w	FALSE	FALSE	failed	1.4E+09	1	{"web"
39	962	Warzone	{"id":34,"r	248173	GB	1.4E+09	{"id":5132	GBP	Â£	FALSE	USD	1.4E+09	FALSE		1.26111	35000	1.9E+09		FALSE		1.4E+09	{"id":3922	Mutant Chronicles V		{"key":"as	161852	{"id":4683	mutant-cl	https://w	TRUE	FALSE	successfu	1.4E+09	1.56828	{"web"
40	63	Creating a	{"id":54,"r	1665	US	1.5E+09	{"id":5497	USD	$	TRUE	USD	1.5E+09	FALSE		1	750	1.1E+09		FALSE		1.5E+09	{"id":2449	Kawaiiju: Cute Kaiju		{"key":"as	1665	{"id":3294	kawaiiju-	https://w	TRUE	FALSE	successfu	1.5E+09	1	{"web"
41	135	Coding Ur	{"id":334,	8026	DE	1.5E+09	{"id":5772	EUR	â‚¬	FALSE	USD	1.5E+09	FALSE		1.13651	3300	1.1E+09		FALSE		1.5E+09	{"id":6499	Coding Unicorn Shie		{"key":"as	6759	{"id":3059	coding-un	https://w	TRUE	TRUE	successfu	1.5E+09	1.17325	{"web"
42	0	The Ambe	{"id":298,	0	US	1.4E+09	{"id":1562	USD	$	TRUE	USD	1.4E+09	FALSE		1	5000	1.1E+09		FALSE		1.4E+09	{"id":2414	The Ambercrest Ma		{"key":"as	0	{"id":2050	the-ambe	https://w	FALSE	FALSE	failed	1.4E+09	1	{"web"
43	0	Through n	{"id":54,"r	0	GB	1.4E+09	{"id":1209	GBP	Â£	FALSE	USD	1.4E+09	FALSE		1.25075	1500	9.2E+08		FALSE		1.4E+09	{"id":3897	Awareness of our be		{"key":"as	0	{"id":1345	awarenes	https://w	FALSE	FALSE	failed	1.4E+09	1.60887	{"web"
44	17	I'm makin	{"id":293,	411	US	1.5E+09	{"id":6356	USD	$	TRUE	USD	1.5E+09	FALSE		1	500	1.1E+09		TRUE		1.5E+09	{"id":2489	They're Playing Our S		{"key":"as	411	{"id":3515	theyre-pla	https://w	FALSE	FALSE	live	1.5E+09	1	{"web"
45	207	Custom d	{"id":34,"r	9199	US	1.4E+09	{"id":1343	USD	$	TRUE	USD	1.4E+09	FALSE		1	4500	5.1E+08		FALSE		1.4E+09	{"id":2383	Thematic Fate/Fudg		{"key":"as	9199.23	{"id":4259	thematic-	https://w	TRUE	FALSE	successfu	1.4E+09	1	{"web"
46	1	We are pr	{"id":54,"r	20	US	1.3E+09	{"id":1833	USD	$	TRUE	USD	1.3E+09	FALSE		1	5000	7.7E+08		FALSE		1.3E+09	{"id":2450	Yellow Fever Love In		{"key":"as	20	{"id":5989	yellow-fe	https://w	FALSE	FALSE	failed	1.3E+09	1	{"web"
47	7	An image,	{"id":301,	306	CA	1.5E+09	{"id":2983	CAD	$	TRUE	USD	1.5E+09	FALSE		0.74859	50000	1.5E+09		FALSE		1.5E+09	{"id":4063	The Myth of Gaia - O		{"key":"as	395	{"id":3163	the-myth-	https://w	FALSE	FALSE	failed	1.5E+09	0.80147	{"web"
48	43	A rich, bla	{"id":250,	1298	US	1.5E+09	{"id":3660	USD	$	TRUE	USD	1.5E+09	FALSE		1	1000	1.1E+09		FALSE		1.5E+09	{"id":2443	Rebirth of the Gangs		{"key":"as	1298	{"id":2384	rebirth-of	https://w	TRUE	FALSE	successfu	1.5E+09	1	{"web"
49	153	A double	{"id":34,"r	5569	US	1.4E+09	{"id":2031	USD	$	TRUE	USD	1.4E+09	FALSE		1	1000	7.9E+08		FALSE		1.4E+09	{"id":2473	Free Trader / Grend		{"key":"as	5569	{"id":4571	free-trade	https://w	TRUE	FALSE	successfu	1.4E+09	1	{"web"
50	19	Fashion P	{"id":278,	2436	US	1.4E+09	{"id":1708	USD	$	TRUE	USD	1.4E+09	FALSE		1	2000	2.1E+09		FALSE		1.4E+09	{"id":7536	Wandering Fashion		{"key":"as	2436	{"id":1099	wanderin	https://w	TRUE	FALSE	successfu	1.4E+09	1	{"web"
51	166	Easy, Sim	{"id":334,	21189	JP	1.5E+09	{"id":2099	JPY	Â¥	FALSE	USD	1.5E+09	FALSE		0.00881	2000000	5.7E+08		FALSE		1.5E+09	{"id":1118	DIY Cardboard Musi		{"key":"as	2219283	{"id":3280	diy-cardb	https://w	TRUE	TRUE	successfu	1.5E+09	0.00903	{"web"

DATA ACQUISITION

- Challenges around data storage & training – testing data creation
  - Initial data set size: 30.6MB (Pandas data frame offered terrible performance)
  - Campaigns were duplicated across multiple scrapes (different snapshots) – needed to specifically isolate instances of records posted within 15 days of campaign
    - Needed to avoid data leakage regarding core metrics like backers, amount pledged, etc. and avg campaign was 30 days long.
  - Additional data processing needed to happen, especially the unix formatted datetime columns

- Solution
  - Created local instance of SQLite db on external drive
  - Create script to load files row by row into master table
  - Query master table for specific subsets of data
  - Merge to create data set for applying NLP and modeling



Create instance of SQLite db ‘Kickstarter’ on external hd

Define master table used

Walk through all csv files in different folders

Read & load campaigns row-by-row into master table

Query master table for relevant datasets

DATA PROCESSING & MERGING

- Two data sets needed:
  - Projects Outcomes: unique list of projects with end state – the labels (‘State’)
  - Projects Starting: snapshot of campaigns in the first 15 days of launching
    - Features used: Name, Blurb, Creator, Category, Goal, Created At, Launched At, Deadline, Location
  - Merged dataset – very well balanced with regard to class:
    - Failed: 823 campaigns
    - Successful: 796 campaigns

- Additional Processing Required
  - Unpacking nested JSON like columns: creator, category, location
  - NLP Specific:
    - Leveraged Spacy by:
      - Using pre-trained statistical model for English
      - Removing stop words, performing lemmatization, and converting tokens to lower case
    - After processing, text columns were then combined into a single column to then be used for BAG of Words, N-grams and TFIDF later

FEATURE ENGINEERING & SELECTION

- Given limited sample size & available features, feature selection options were limited & I decided to use as many as possible

- Aside from text features, train-test included:
  - Goal, category, country, state, city, time to launch, launch to deadline, launched month, launched year, created month, created year, deadline month, deadline year (total: 13)

- Process:
  - Pipelines created to select numeric, text, & categorical columns
  - Numeric pipeline:
    - Utilized SimpleImputer to fill nan's with 0's (instead of the mean, etc) as using the mean would have been misleading (i.e. some campaigns could truly have taken less time to launch, etc.)
  - Categorical pipeline:
    - Utilized SimpleImputer, OneHotEncoder
  - Text pipeline:
    - This is where I experimented with different combinations of NLP techniques including Bag of Words, TFIDF Vectorizer, N-grams to explore which techniques could work best

MODEL SELECTION & PERFORMANCE

I'M A BACKER

- Because of the inclusion of text data & feature engineering techniques like Bag of Words, N-grams, etc. + the small sample size, a big driver of model selection was focused on the ability to utilize a sparse array

- Classification models known to do well with sparse arrays & high dimensional data (even in cases where columns outnumbered actual record size) included:
  - Logistic Regression
  - Linear SVC (SVM's in general)
  - Random Forest Classifier

- Results:
  - Models performed equally well around ~65% accuracy (with cross validation (cv=5))
  - Hyperparameter tuning of Random Forest model with bi-grams & TFIDF transformer added additional ~2% lift in model accuracy

- Interpretation:
  - Given the near equally dismal performance of all model variants, additional data + text based features could potentially have provided more insight into drivers of campaign success

MODEL SELECTION & PERFORMANCE

<table>
<tr><th>Model Type</th><th>Bag of Words</th><th>TFIDF</th><th>N-grams</th><th>Hyperparameter Tuning</th></tr>
<tr><td rowspan="2">Logistic Regression</td><td>Model Variant 1:
Accuracy: 65%</td><td></td><td></td><td></td></tr>
<tr><td colspan="2">Model Variant 3:
Accuracy: 66%</td><td></td><td></td></tr>
<tr><td>Linear SVC</td><td>Model Variant 2:
Accuracy: 64%</td><td></td><td></td><td></td></tr>
<tr><td>Random Forest Classifier</td><td></td><td colspan="2">Model Variant 4:
Accuracy: 66%</td><td>Hyperparamter tuning:
Accuracy: 68%</td></tr>
</table>

TAKE AWAY & NEXT STEPS

Project Specific Mentions:

- For NLP to be effective, lots of data is needed + for NLP to be effective in prediction, as many or more labeled samples
- Time duration + frequency of data collection key in being able to (1) isolate early conditions of observatory units + (2) understand time-based trends like pledging behavior.
- Model interpretability key to socializing model outcomes + decisions; pipelines offer way to streamline & productionalize models but also make interpretability challenging

Next Steps

- Applying learnings from this project to similar projects with similar desired outcomes + higher quality, volume data appropriate for NLP
- Exploring the use of Deep Learning models involving NLP tasks

ABOUT ME

ABOUT ME

Applied analytics and data science evangelist, Mikiko Bazeley is a seasoned analyst with 5+ years of working in high-impact roles for start-ups and enterprise tech companies.

A UCSD Economics & Anthropology graduate, Mikiko aims to use her experience in social research & modeling to strategically leverage data science in order to drive new insights for sales, marketing, finance & customer success organizations. Mikiko is also GIS & Supply Chain Management certified.

Prior to joining WalkMe (where she focuses on scaling data science & global sales analytics) Mikiko worked as a Data Scientist at Autodesk (focused on understanding product adoption & user health), as well as assisting with scaling strategic finance initiatives at Sunrun (the largest residential solar company in the US).

Please feel free to reach out:

- LinkedIn: https://www.linkedin.com/in/mikikobazeley/
- Email: mmbazel (at) gmail (dot) com