DESCRIBING & PREDICTING THE SUCCESS RATE OF KICKSTARTER PROJECTS

David Forteguerre, Ziwei Pi, Haoxuan Shi

1. PROJECT OVERVIEW AND OBJECTIVES

Online fundraising platforms are becoming more and more popular. However, the majority of projects turn out to be ***unsuccessful***

Goals

- Use ***descriptive statistics*** to learn more about already completed projects
- Use ***machine learning*** to find the factors that play a key role in determining the outcome of a project (i.e. success or failure) and make predictions

2. DATA DESCRIPTION

Data source: Kickstarter Projects dataset, available on Kaggle

		name	category	main_categor	currency	country	status	launched_date	deadline	days_allotted	fundraising_goal	fundraising_pledged	completion_percentage	backers
2	0	The Songs of	Poetry	Publishing	GBP	GB	failed	8/11/15	10/9/15	59	1533	0	0	0
3	1	Greeting From	Narrative Film	Film & Video	USD	US	failed	9/2/17	11/1/17	60	30000	2421	8	15
4	2	Where is Hanl	Narrative Film	Film & Video	USD	US	failed	1/12/13	2/26/13	45	45000	220	0	3
5	3	ToshiCapital F	Music	Music	USD	US	failed	3/17/12	4/16/12	30	5000	1	0	1
6	4	Monarch Espr	Restaurants	Food	USD	US	successful	2/26/16	4/1/16	35	50000	52375	104	224

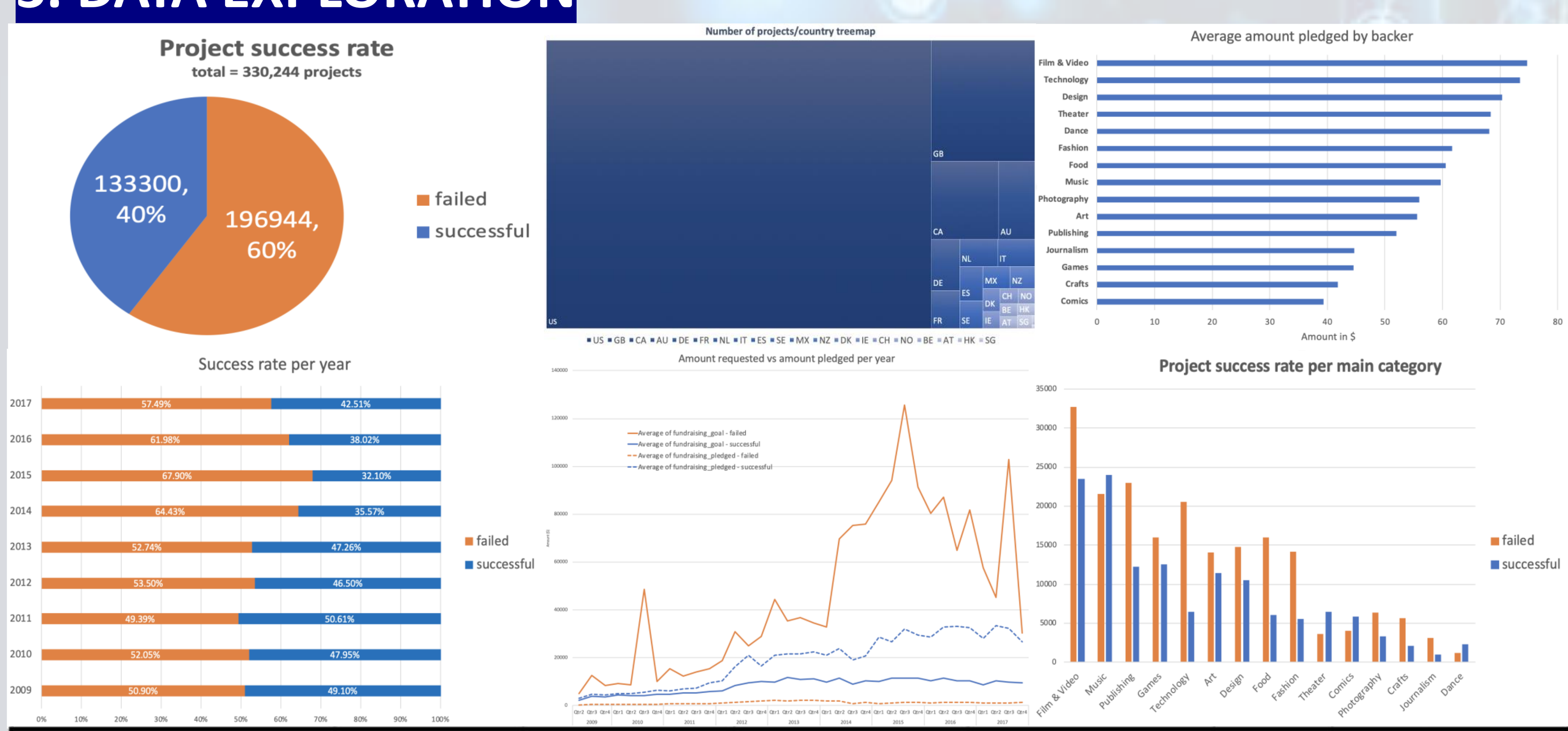
(cleaned dataset overview)

Data cleaning:

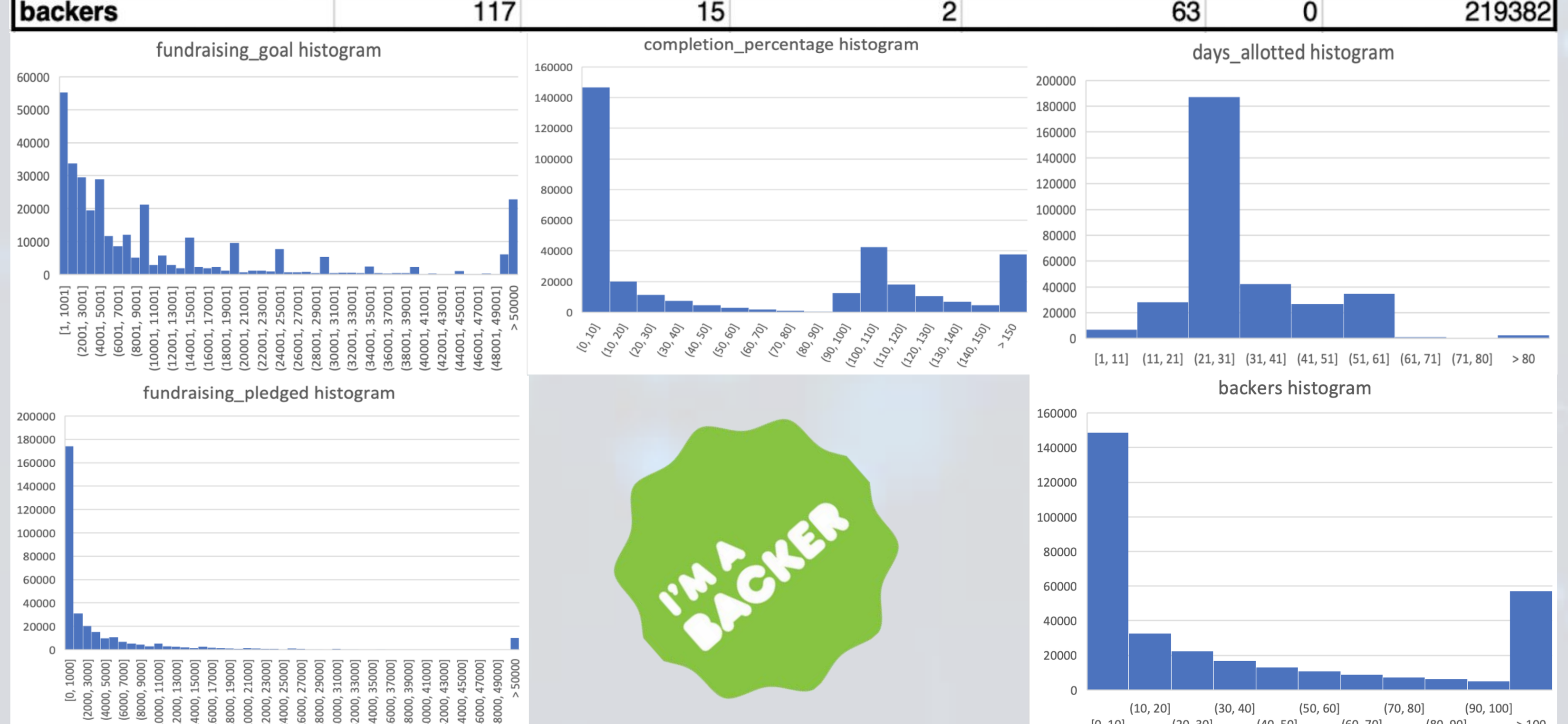
- Renamed and reordered columns
- Fixed data types
- Parsed variables with dates
- Only kept projects that *failed* or were *successful* *(88% of original data)*
- Removed country named 'N,0""'
- Dropped redundant columns/incorrect data
- Removed rows that had a fundraising goal of $0
- Created columns (*day_allotted, completion_%)*
- Checked for NAs

CLEANING	*before*	*after*
# rows	377,364	**330,244**
# columns	15	13

3. DATA EXPLORATION



SUMMARY STATISTICS	mean	median	Q1	Q3	min	max
fundraising_goal	$ 41,616.83	$ 5,000.00	$ 2,000.00	$ 15,000.00	$ 1.00	$ 166,361,390.00
fundraising_pledged	$ 9,962.68	$ 787.00	$ 50.00	$ 4,614.00	$ -	$ 20,338,986.00
days_allotted	34	30	30	36	1	92
completion_percentage	344	20	0	109	0	10427700
backers	117	15	2	63	0	219382



4. MODEL DESCRIPTION

Classification task with **a binary** outcome (success, failure)

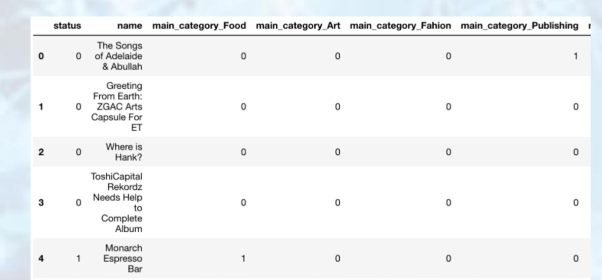
Features used for machine learning

- **1st round of models:** *main_category, country, launch_month, launch_day, launch_weekday, days_alloted, fundraising_goal*
- **2nd round of models:** *main_category, **name**, country, launch_month, launch_day, launch_weekday, days_alloted, fundraising_goal*

Feature engineering

- *main_category, country, launch_month, launch_day, launch_weekday* → dummy variables
- *days_allotted* → untouched
- *fundraising_goal* → log transformation *(common with positive right-skewed data w/ outliers)*
- *name…*
  - ✓ → lowercased, punctuation removed with regex, tokenized
  - ✓ → stopwords removed *(list taken from tm package in R)*
  - ✓ → vectorized, *converted to* term frequency–inverse document frequency *(tfidf)*

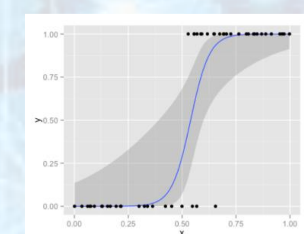
*Number of features without text: **56***
*Number of features with text: **56 + 92,680 tokens***

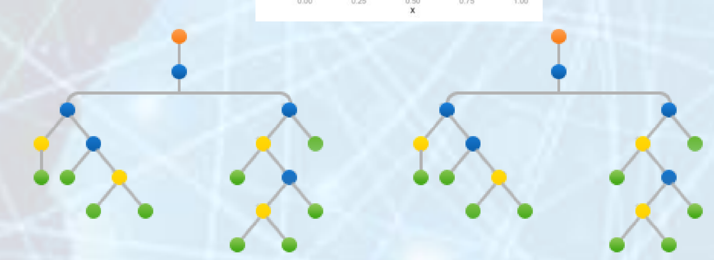


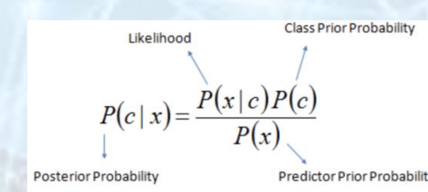
(dataset after feature engineering)

Models used in round #1 (*without* text features)

- **Logistic regression** *(scaled data)*
  - $\lambda = 0$ $\alpha = 0$ *(no elastic net regularization)* ✓
  - $\lambda = 0.02$ $\alpha = 0.2$
  - $\lambda = 0.1$ $\alpha = 0.4$
- **Random forest** *(unscaled data)*
  - maxDepth=1 numTrees=60
  - maxDepth=6 numTrees=100
  - maxDepth=6 numTrees=80
  - maxDepth=15 numTrees=80 ✓
- **Naïve Bayes**
  - smoothing=1.0 *on unscaled data* ✓
  - smoothing=5.0 *on unscaled data*
  - smoothing=1.0 *on scaled data*







Models used in round #2 (*with* text features)

- **Logistic regression**
  - *With same parameters as found for the best model in round #1 for maximizing run time*
- **Random forest**
  - *Eventually dropped as a candidate algorithm for round #2 due to the large number of text features (we conducted many experiments – the server would always crash after a few hours)*
- **Naïve Bayes**
  - *With same parameters as found for the best model in round #1 for maximizing run time*

5. MODEL COMPARISON METRICS

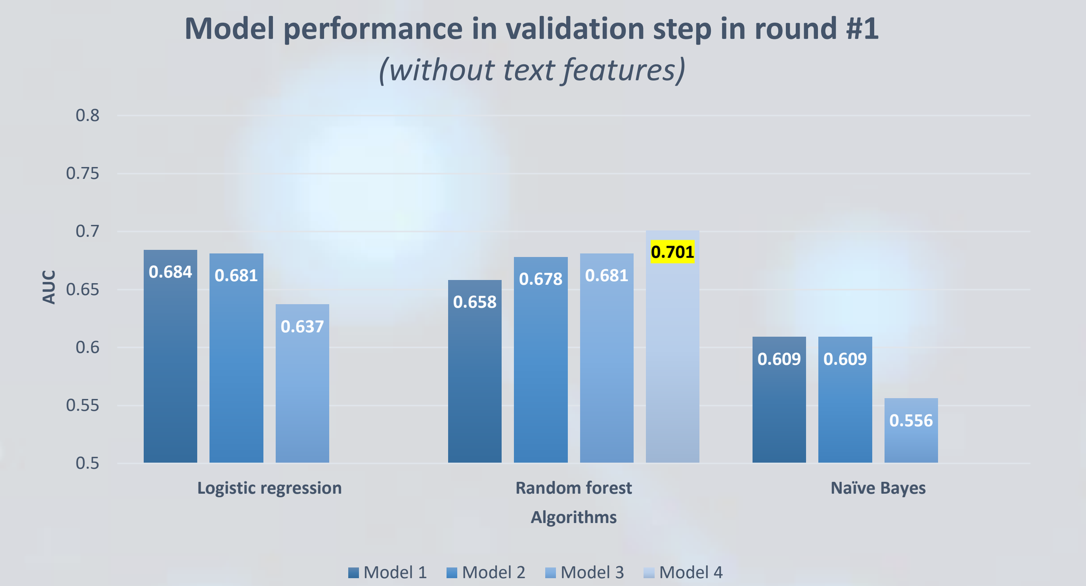
The data was split into **3** datasets:

training (60%) *used to fit the models* → **validation (30%)** *used to pick best model of the same algorithm* → **testing (10%)** *used to compute generalization performance of best model on unseen data*

AUC ROC curve was used for evaluating the **performance** of models

6. RESULTS – PREDICTION PERFORMANCE

1. validation performance:



2. testing performance:

ROC AUC (testing data) of models	*without* text features	*with* text features
Logistic regression	0.685	0.657
Random forest	**0.701** [highest accuracy]	*[candidate dropped]*
Naïve Bayes	0.612	0.460

Confusion matrix	pred(0)	pred(1)
actual(0)	16,825 TN	2,975 FP
actual(1)	8,365 FN	5,049 TP
precision: 0.63	*recall 0.38*	

7. RESULTS - INFERENCE

Logistic regression feature importance
*In round #1 without text features (**AUC=0.685**)*

feature	coefficient
main_category_Dance	0.78013493
main_category_Theater	0.73181417
main_category_Comics	0.53967011
country_HK	0.38315105
…	…
main_category_Journalism	-0.948189
main_category_Crafts	-1.0626787
days_allotted	-1.4184967
log_fundraising_goal	-5.1330291

Random forest feature importance
*In round #1 without text features (**AUC=0.703**)*

feature	importance
log_fundraising_goal	0.35976
days_allotted	0.1956476
main_category_Technology	0.04770462
main_category_Music	0.04627876
main_category_Theater	0.03969938
main_category_Fahion	0.03245688
main_category_Comics	0.02516374
main_category_Food	0.02272381
…	…

8. CONCLUSION

- Performance was quite **high** overall with random forest (approx. 70%)
- As a result, success rate may not only depend on the data and could also greatly vary based on factors that are unavailable to us/specific to each project
- Text features did not help improve performance with these models

FUTURE WORK

- Rerun logistic regression on text features using elastic net regularization despite the very high computational cost, and utilize/implement more algorithms which allow for more complex learning (e.g. neural networks)
- Attempt to predict *fundraising_pledged* and *number of backers (i.e. numeric variables with a linear relationship instead of categorical ones/classification outcomes)*
- Retrieve more data/projects to improve performance

School of Information Studies
Syracuse University