



Predicting Campaign Success

Brian McMahon

21 February 2018

What is Kickstarter?

Leading crowdfunding platform,
raising small amounts of money
from large amounts of people

Since 2009:

- \$3.5 billion raised
- 139,000+ projects funded
- 14 million project backers

*“Our mission is to help
bring creative projects to life”*



\$2,533 pledged
12% funded
29 days to go

[Graphic Novels](#)

Example Campaign

Model Overview

— — —

- Dataset scraped from Kickstarter*
- From January 2016 to today
- 8 -> ~160 features (using “one-hot” dummy variables)
- Analysis:
 - Initial screen of several models
 - *GridSearchCV to optimize parameters*
 - Deep dive on two high performers

* Kickstarter scraped dataset courtesy of webrobots.io.

Dataset Features

— — —

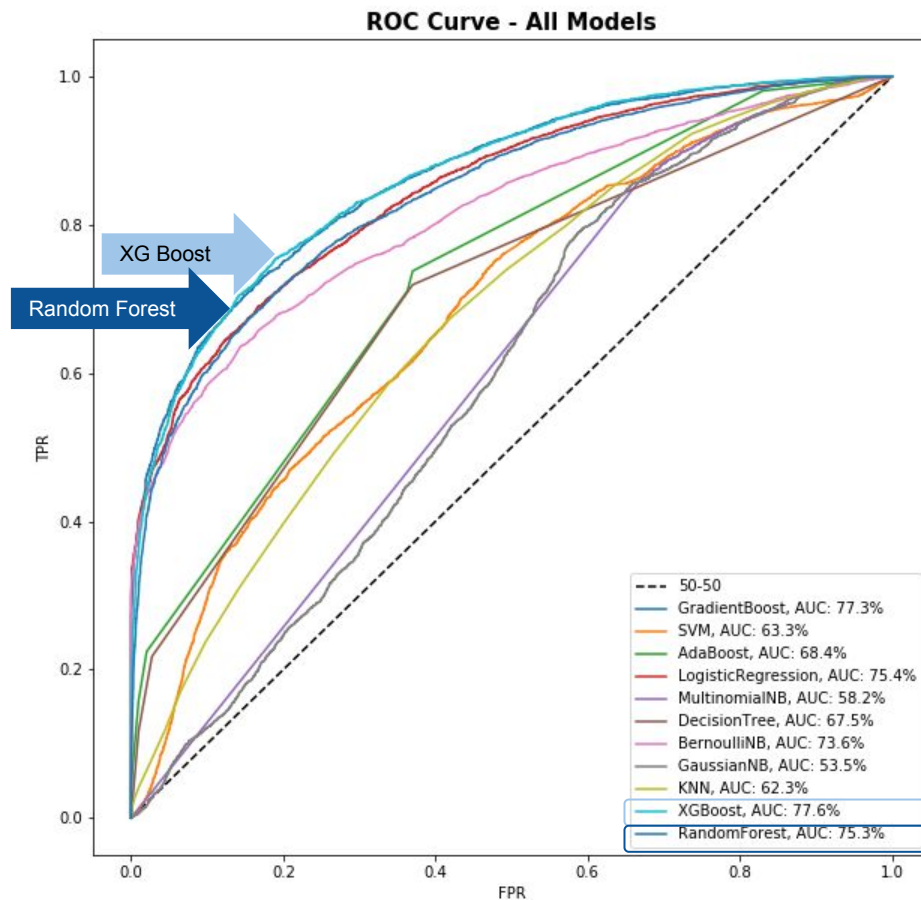
X Features
Category
Subcategory
Staff Pick
Goal (US\$)
Country
Currency
Campaign Length (Days)
Blurb Length (Words)

Y Targets
Backers Count
Pledged (US\$)
State (“Success”, “Fail”)

*Categorical; our
supervised ML model
response*

ROC Curve: All Models

	accuracy	auc	f1_f	f1_s	precision_f	precision_s	recall_f	recall_s
XGBoost	77.6%	77.6%	75.9%	79.1%	73.9%	81.0%	78.0%	77.3%
GradientBoost	77.2%	77.3%	75.6%	78.7%	73.3%	80.9%	78.0%	76.6%
LogisticRegression	75.3%	75.4%	73.5%	76.9%	71.3%	79.0%	75.9%	74.9%
RandomForest	75.5%	75.3%	73.0%	77.5%	72.6%	77.9%	73.4%	77.2%
BernoulliNB	73.3%	73.6%	72.1%	74.3%	68.1%	78.6%	76.7%	70.5%
AdaBoost	68.9%	68.4%	64.7%	72.2%	66.4%	70.8%	63.0%	73.8%
DecisionTree	67.9%	67.5%	63.9%	71.1%	64.9%	70.3%	63.0%	71.9%
SVM	64.7%	63.3%	55.8%	70.6%	64.1%	65.0%	49.4%	77.2%
KNN	63.4%	62.3%	55.5%	69.0%	61.6%	64.5%	50.6%	74.0%
MultinomialNB	61.4%	58.2%	37.4%	72.1%	69.7%	59.7%	25.5%	90.9%
GaussianNB	57.9%	53.5%	15.4%	72.0%	82.5%	56.7%	8.5%	98.5%



Business Case: What we care about

— — —

Campaign Creator

Cares about: **Precision**

(low False Positive, Type I Error)

- Decrease risk of creator's campaign in fact failing (*when they believed it would succeed*)
- Cost: campaign investment

Campaign Backer

Cares about: **Recall**

(low False Negative, Type II Error)

- Decrease risk of backer missing a success
- Cost: missed success

***∴ Maximize both precision and recall
(with F1 Score)***

Random Forests

accuracy ✓ auc ✓ f1_f ✓ f1_s ✓ precision_f ✓ precision_s ✓ recall_f ✓ recall_s ✓

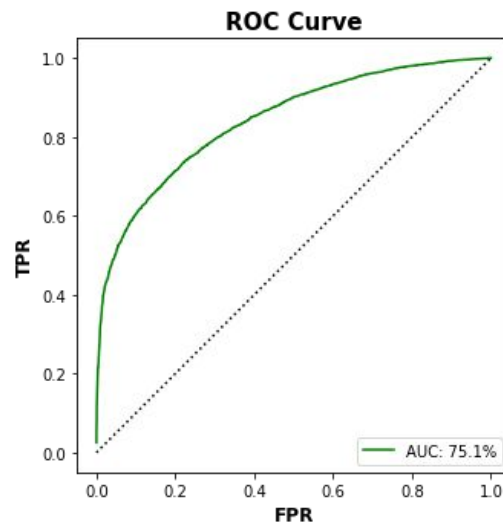
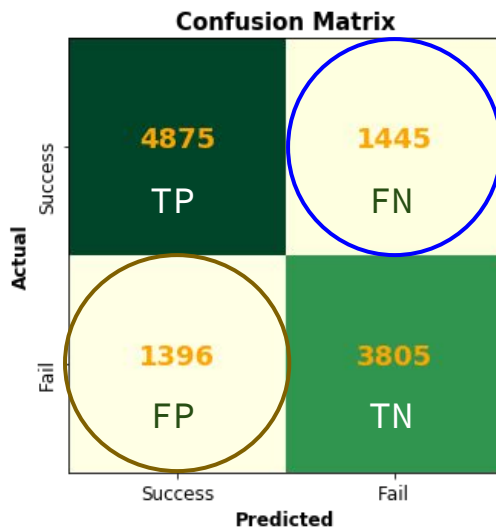
RandomForest 75.4% 75.2% 72.9% 77.4% 72.4% 77.8% 73.3% 77.1%

High Precision **Creator** 👍

High Recall **Backer** 👍

High F1 ✓

High AUC ✓



Random Forests: Feature Importance

importances	features	
0.243471	usd_goal	Keys to campaign success
0.154323	blurb_length	
0.119610	campaign_length	
0.073467	staff_pick	
0.015714	category_name_Apparel	Category matters!
0.013365	category_name_Children's Books	
0.012575	category_name_Video Games	
0.012474	category_main_food	
0.009060	category_name_Nonfiction	
0.008936	category_name_Illustration	

XG Boost





— — —

High Precision **Creator** 

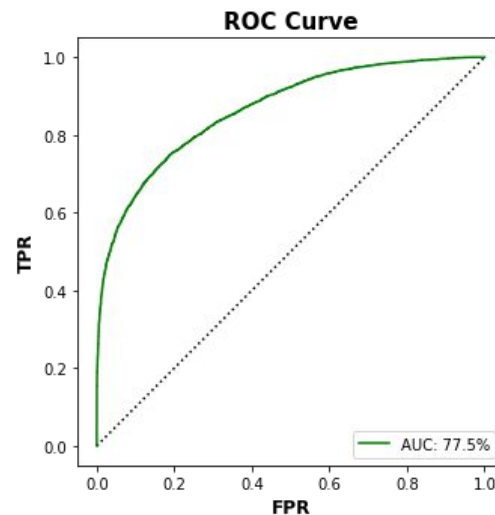
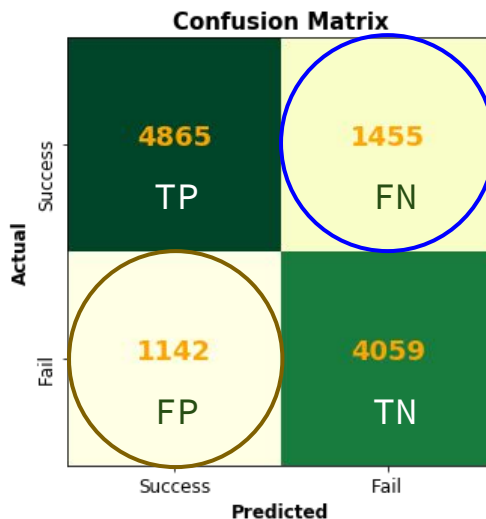
High Recall **Backer** 

High F1 

High AUC 

accuracy  auc  f1_f f1_s precision_f  precision_s recall_f recall_s 

XGBoost	77.5%	77.5%	75.8%	78.9%	73.6%	81.0%	78.0%	77.0%
---------	-------	-------	-------	-------	-------	-------	-------	-------



Feature Exploration

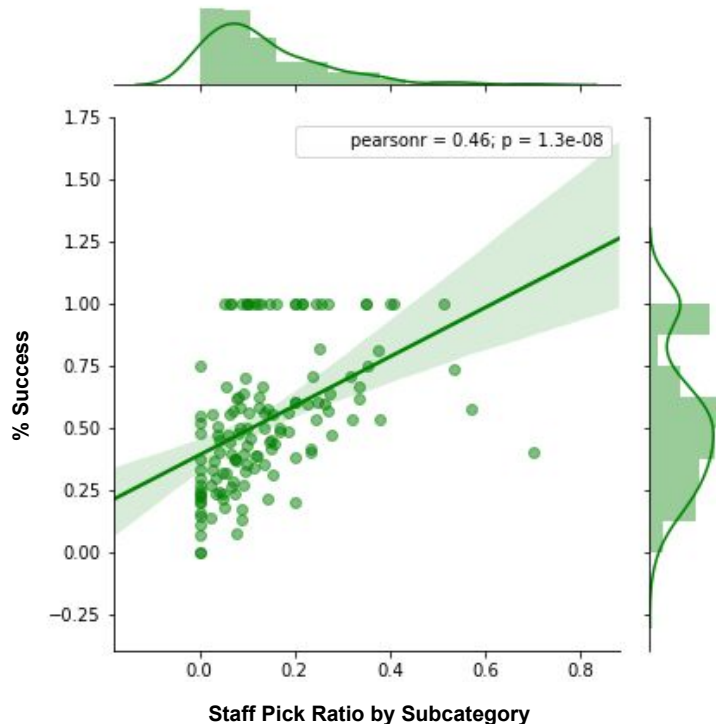
160 subcategories plotted to determine whether a recommendation by kickstarter staff (“staff pick”) affects the success of a campaign*

- x axis: staff picks as % of subcategory total
- y axis: success rate as % of subcategory total

Click on chart to visualize via Flask / D3 web app...

** data from July 2017 to today*

Staff Pick vs. Success Rate by Subcategory



Key Takeaways

— — —

Remember!

When planning a Kickstarter Campaign:

1. Set a **low US\$ goal**
2. Be **recommended** by staff
3. Make your **campaign short**
4. **Concise** description
5. Category **matters**



Example Campaign



Mিনny Spoons - Cashew Butter + Energy Bites

By Ashley Prentice

First created

A female owned cashew butter company dedicated to making delicious products with simple, wholesome, and nutrient dense ingredients.

Questions?



6,990

ged of \$12,000 goal

1

kers

s to go

Back this project

Remind me



All or nothing. This project will only be funded if it reaches its goal by Sun, February 25 2018 7:00 PM PST.

Example Campaign

APPENDIX

Agenda

— — —

Intro to Kickstarter

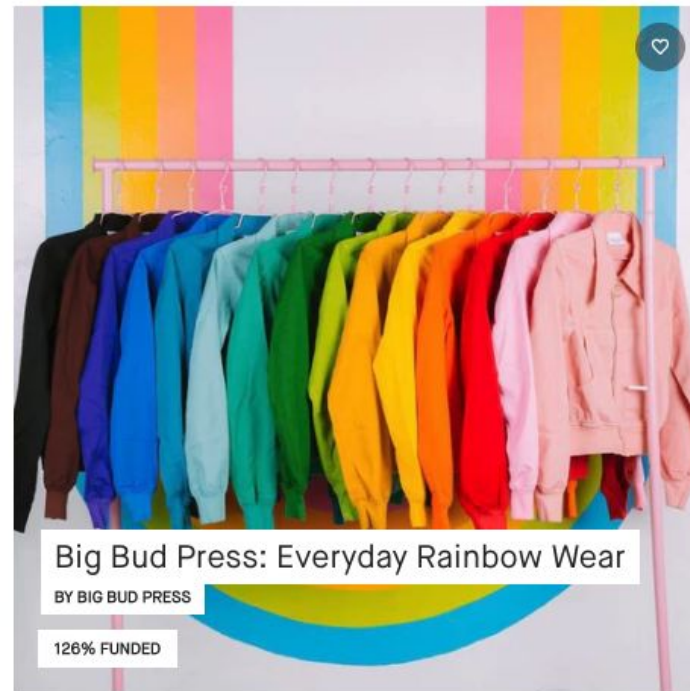
Prediction Model Overview

Data Exploration
with Flask/D3 visualization

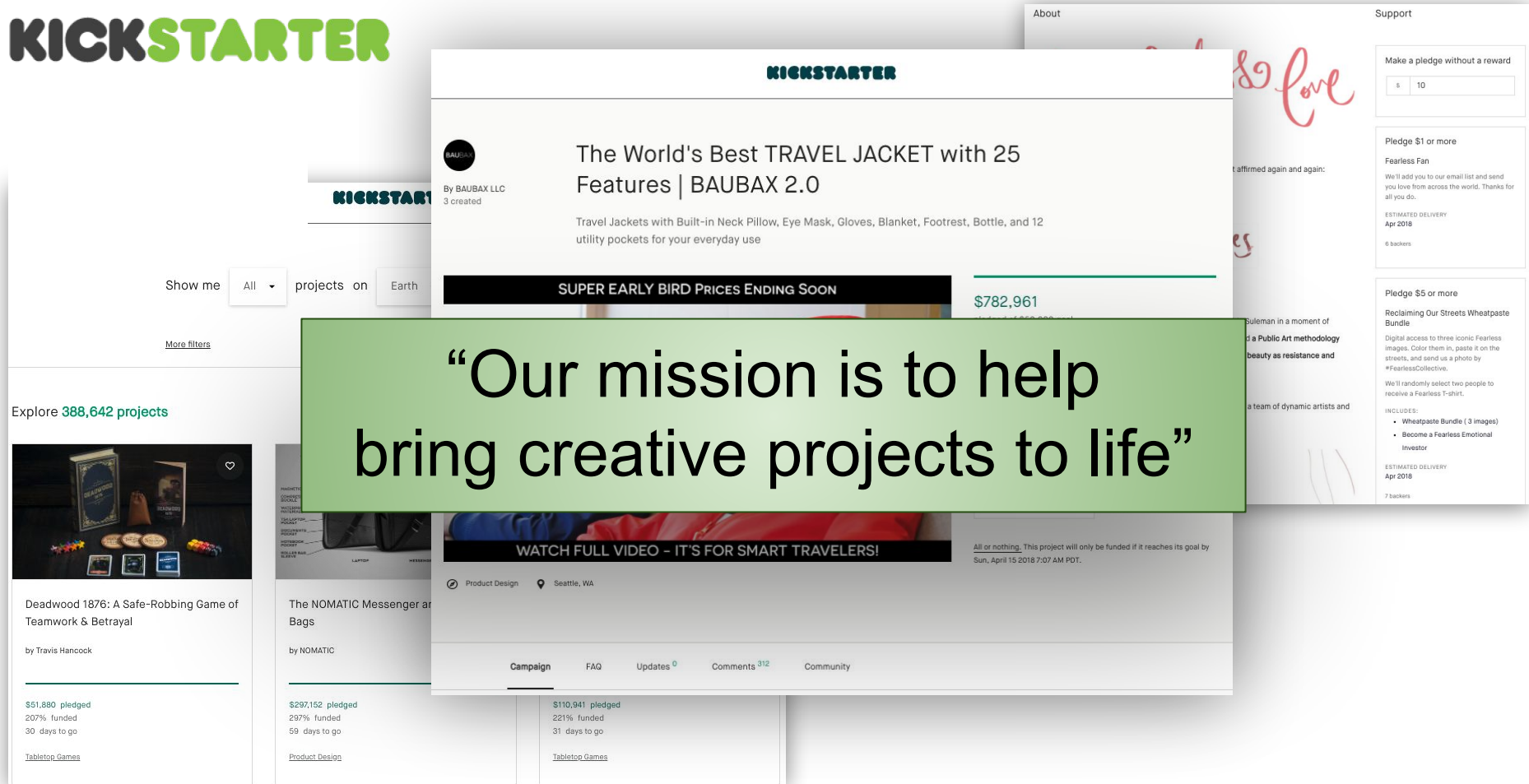
Key Takeaways

Food & Craft [VIEW ALL →](#)

FEATURED PROJECT

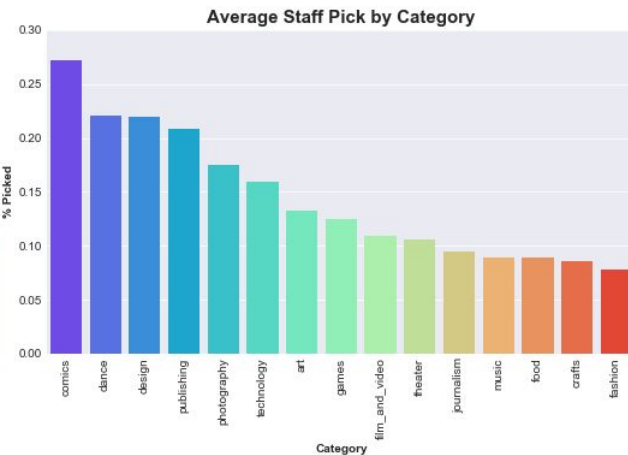
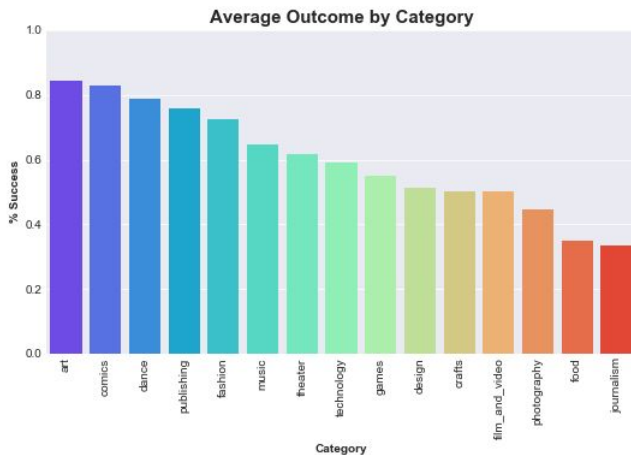
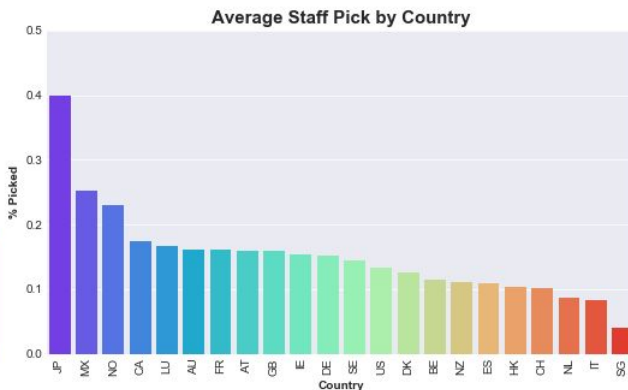
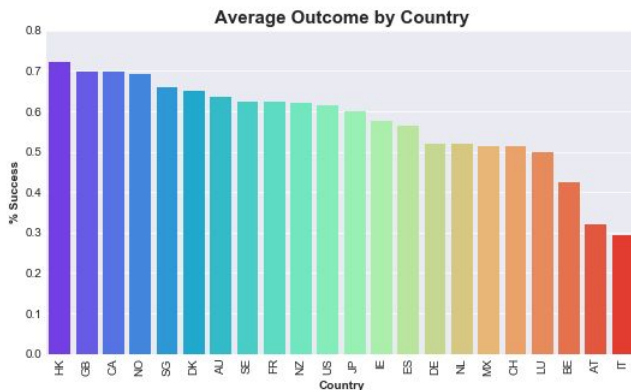


Example Campaign



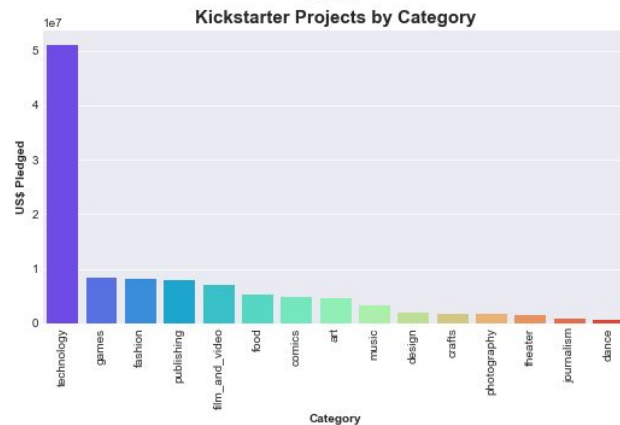
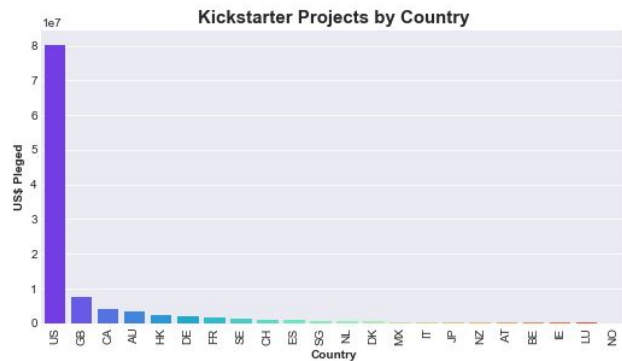
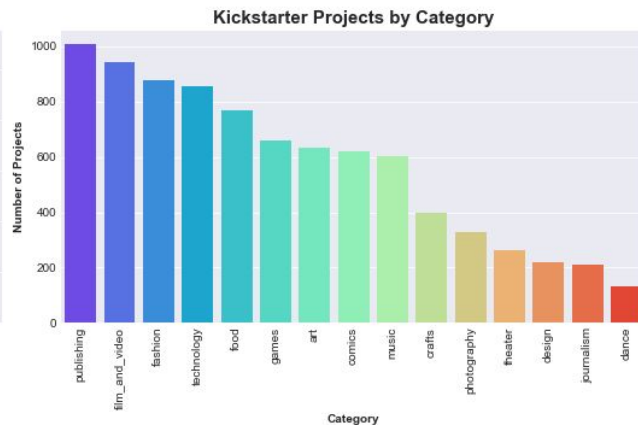
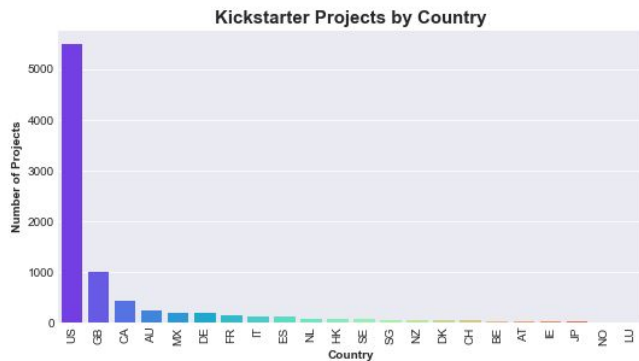
Feature Exploration

— — —



Feature Exploration

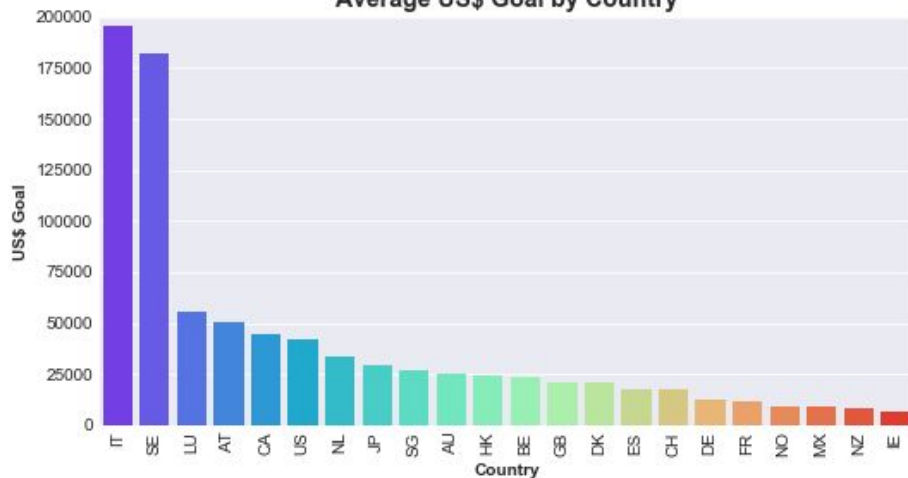
— — —



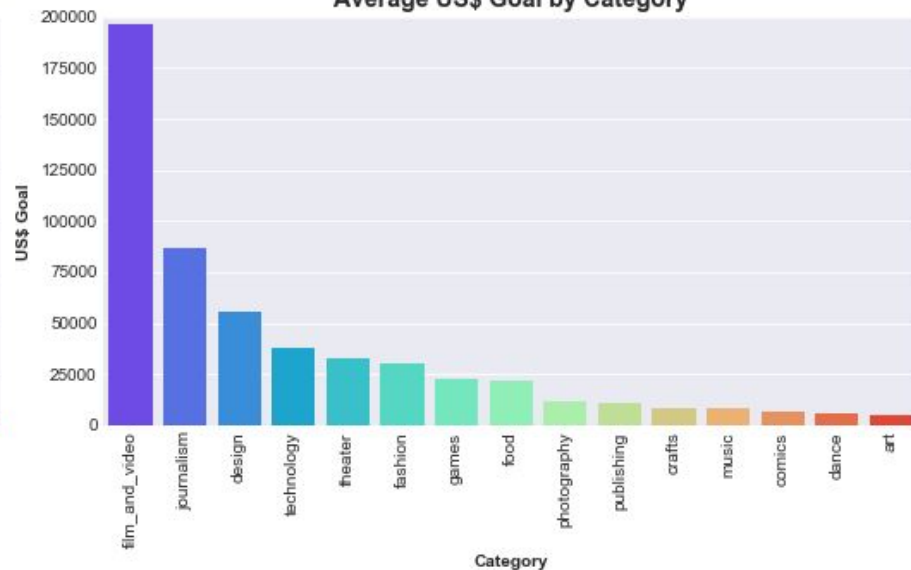
Feature Exploration

Lower goals typically lead to greater success

Average US\$ Goal by Country

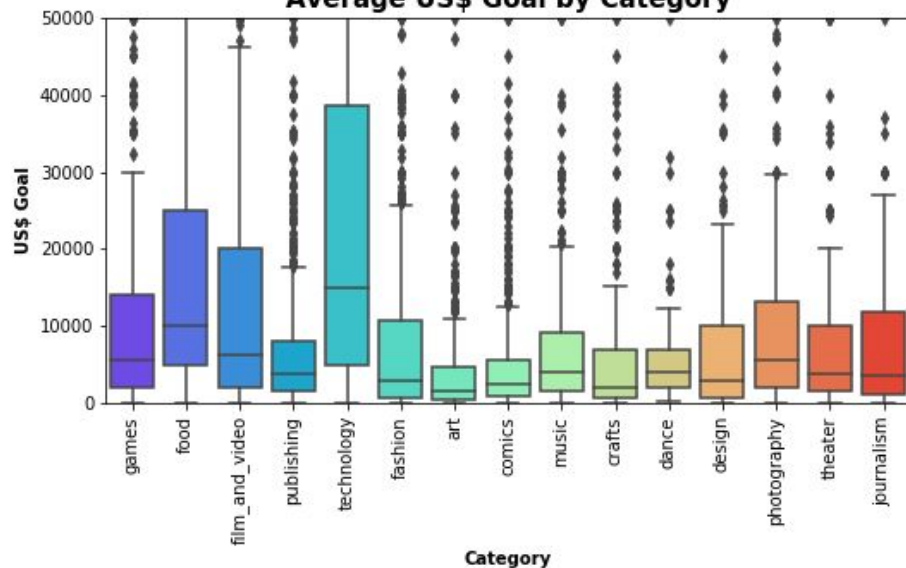


Average US\$ Goal by Category

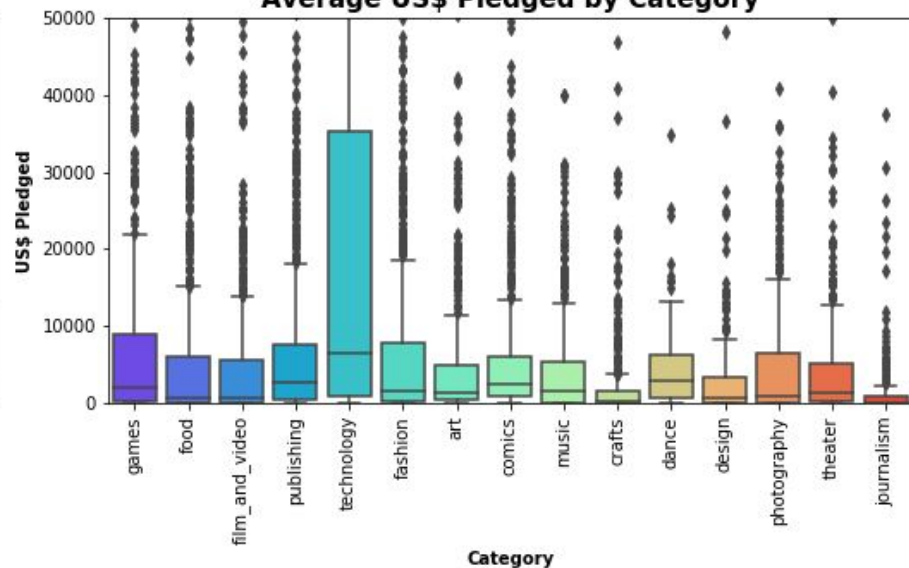


Feature Exploration

Average US\$ Goal by Category



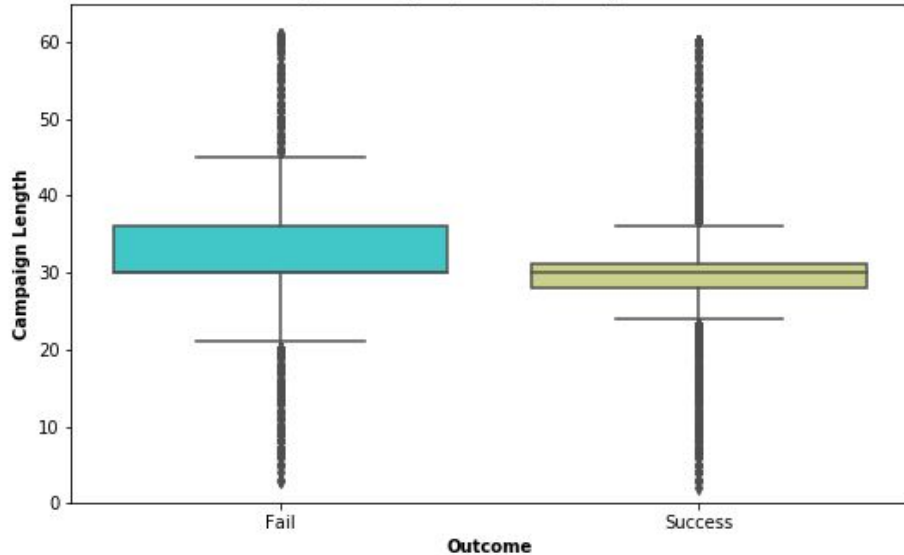
Average US\$ Pledged by Category



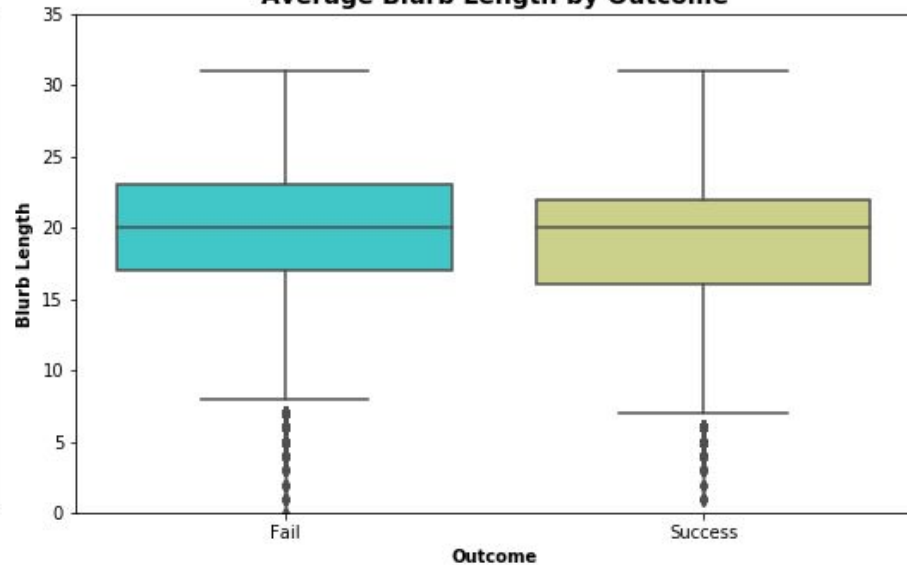
Feature Exploration

— — — *Successful campaigns are, on average, shorter in duration with concise descriptions*

Average Campaign Length by Outcome



Average Blurb Length by Outcome



Model Methodology

— — —

- Dataset scraped from Kickstarter website*
- 160k -> 40k rows, from January 2016 to January 2018
- 8 -> ~160 features (with “one-hot” dummy variables)
- Model refinement procedure:
 - Initial screen, several ML models preliminarily run against data
 - Preprocessing: StandardScaler for SVM, KNN, LogisticRegression
 - *Utilize GridSearchCV to optimize parameters*
 - Deep dive on highest performing models: Random Forests and XG Boost

* Kickstarter scraped dataset courtesy of webrobots.io.

Dataset Features

id	163425	non-null	int64	Unique campaign ID
name	163424	non-null	object	Project Name
state	163425	non-null	object	<i>Response, categorical.</i> Filtered to “Success” or “Fail”
category_main	163425	non-null	object	15 main categories
category_name	163425	non-null	object	~160 subcategories
backers_count	163425	non-null	int64	<i>Response, numerical.</i> # backers
pct_goal_achieved	147802	non-null	float64	<i>Response, numerical.</i> US\$ pledged / US\$ goal
usd_pledged	163425	non-null	float64	<i>Response, numerical.</i> US\$ pledged
usd_goal	147802	non-null	float64	US\$ goal
country	163425	non-null	object	Country
currency	163425	non-null	object	Currency
campaign_length	163425	non-null	int64	Length of campaign (days)
deadline	163425	non-null	object	Project end data
launched	163425	non-null	object	Project launch date
created	163425	non-null	object	Project creation date
staff_pick	163425	non-null	int64	Recommended by Kickstarter staff
creator_name	163425	non-null	object	Name of creator
blurb_length	163425	non-null	int64	Length of intro blurb

Y Value

X Values

Dataset Stats

— — —

	Dataset 1	Dataset 2
Source	Kaggle	WebRobots.io
Initial Shape	(378661, 17)	(192716, 37)
Key features	Category, subcategory, country, currency, goal, campaign length, pledged*, backers*	Category, subcategory, country, currency, goal, staff pick , campaign length, blurb length , pledged*, backers*
Time period	2012 - Present	2016 - Present
* removed from prediction dataset		

Preprocessing

State to Success, fail (dropped live, suspended, cancelled)

StandardScaler for SVM, KNN, Logistic Regression

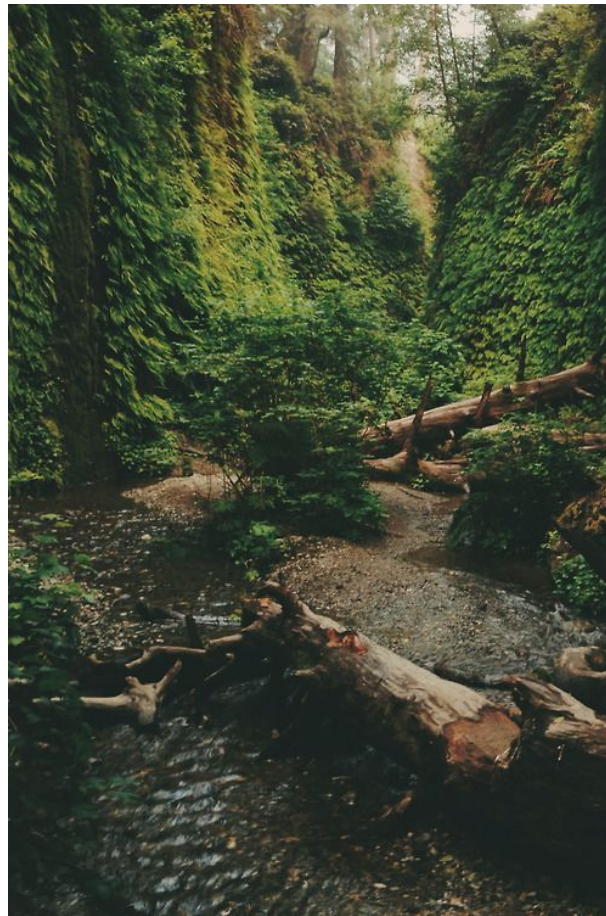
It all starts with SQL...

```
Select * from kickstarter_data2;
```

Random Forests: What it is

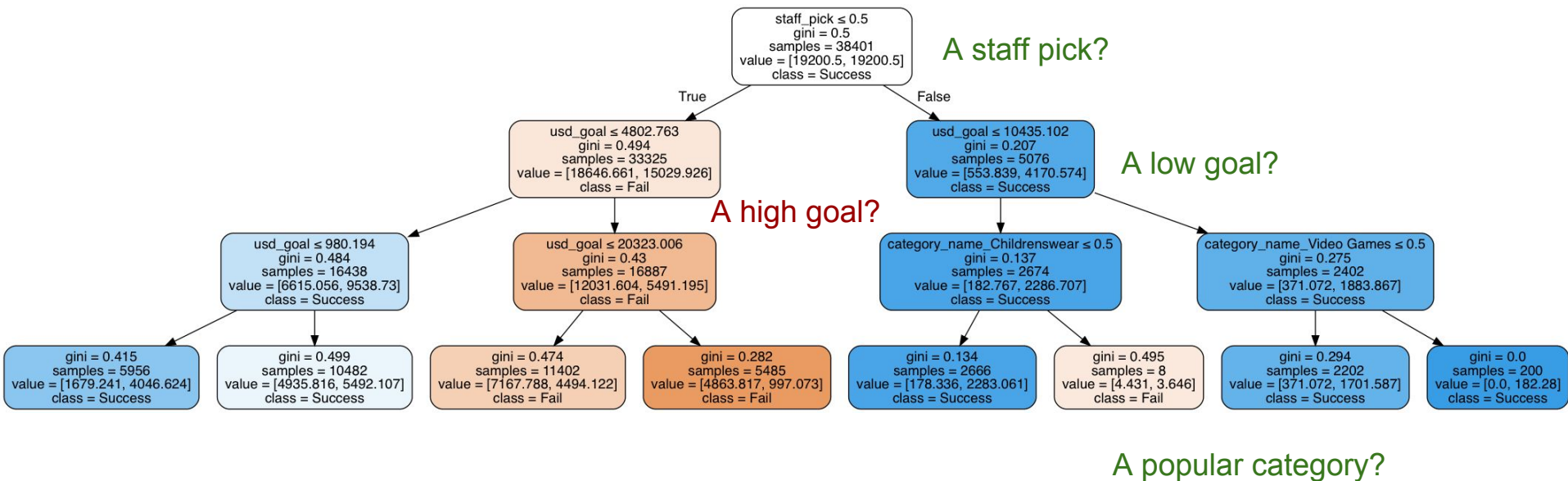
— — —

- Ensemble learning method
- Trains on a multitude of decision trees
- Outputs class that is the mean/mode of individual trees
- Corrects overfitting, a weakness of individual decision trees

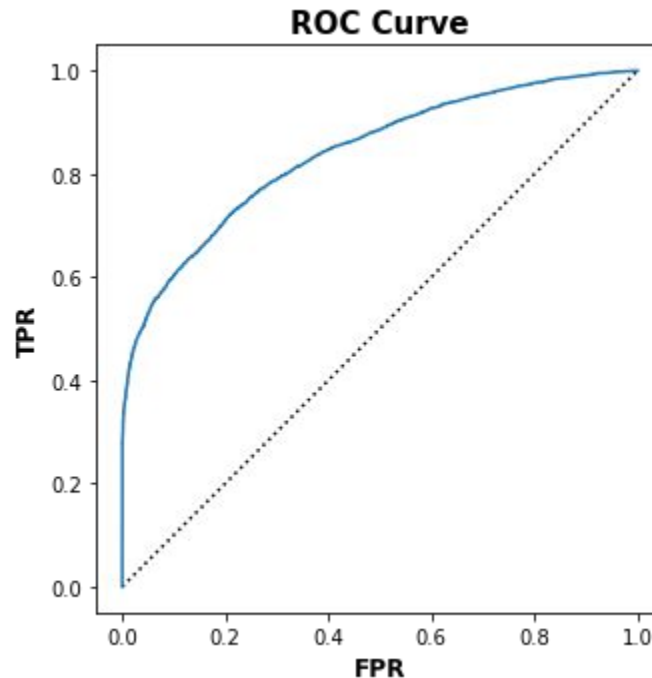
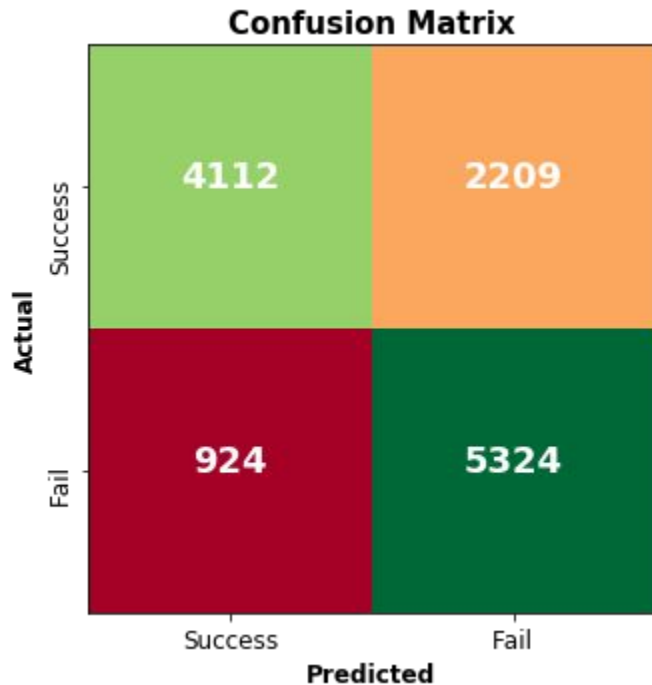


A random forest..

Random Forests: Decision Tree



Gradient Boost



XG Boost: What it is

— — —

- “Extreme Gradient Boosting” is a gradient boosting framework proposed in Friedman’s *“Greedy Function Approximation: A Gradient Boosting Machine”*
- Designed and optimized for boosted tree algorithms
- Popular, highly competitive algorithm for ML competitions

Process / Challenges

— — —

Multiple datasets analyzed - not all datasets come equal!

Had to expand set of features to obtain adequate results

Initial screen with first data set narrowed features to only category, country, currency and goal value

Preprocessing: dummy variables, standardizing, train/test split and cross-validate, removed all but successful and failed (such as live, cancelled)

Initial pass through all models, generating initial ROC curve

Deeper analysis into RF and LR, reasons being [expand]

Next Steps

Integrate twitter analysis of
launched projects

Incorporate numerical US\$
pledge prediction ie with
linear regression

