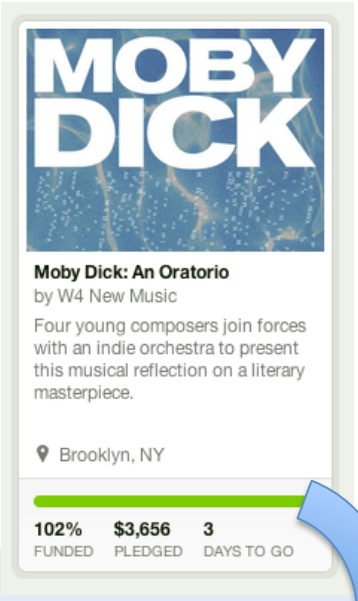
Predicting KICKSTARTER

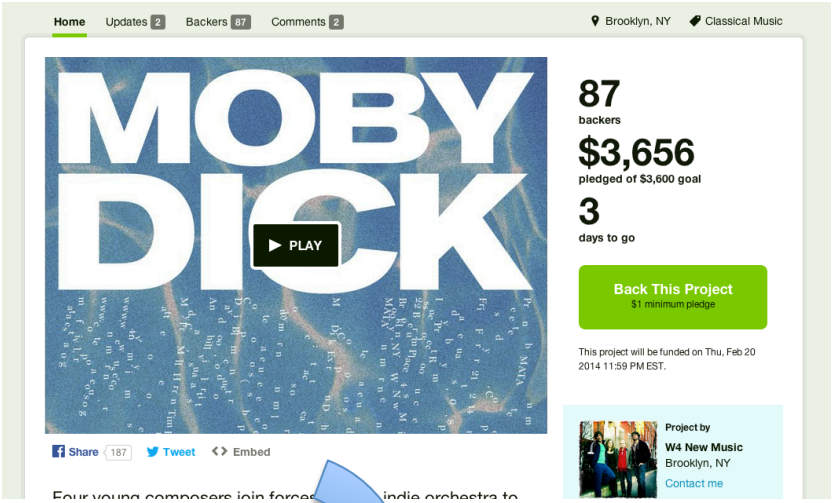
Ruben Naeff

Predicting Success, March 2014
General Assembly Data Science Course

As we speak, a Python script is downloading all Kickstarter projects to a database



Hourly updates to track funding progress



Downloaded once

MongoDB

Widget database
Project ID
Title
Short description
End time
Pledged (in $ or £)
Funding rate
Link to webpage
Location (woeid)
Location name
Time stamp (of download)

Project page database
Project ID
Funding goal
Total pledged (as of download)
Full description (body text)
List of rewards (text and price)
List of images and videos
# of backers
# of comments
# of updates

Project page database
Raw HTML code

120-130 thousand projects
of which 96% closed

Web scraping and database code written by Tobias Tiecke

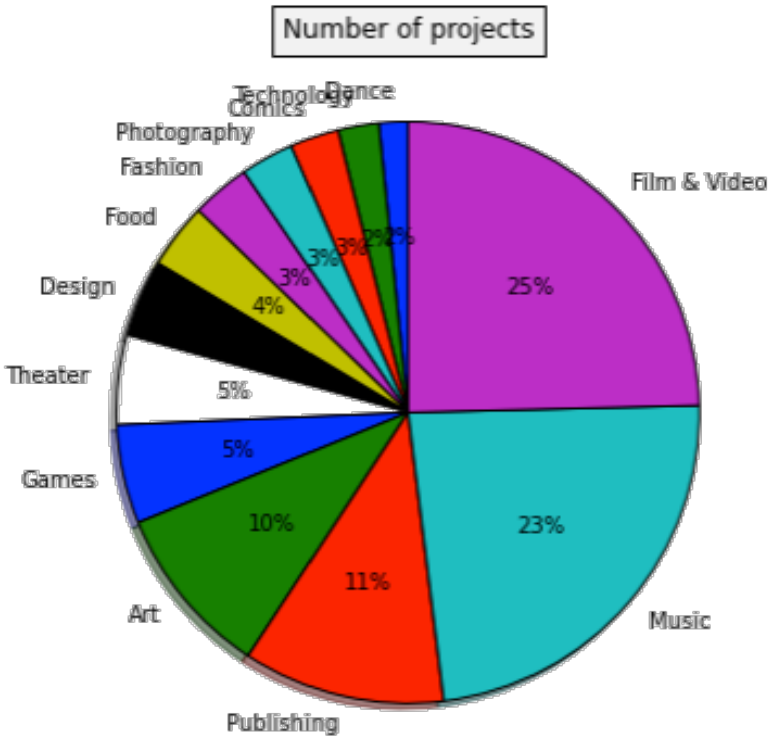
Agenda

Descriptive Statistics

Machine Learning

Film & Video and Music projects make up about half of all Kickstarter projects

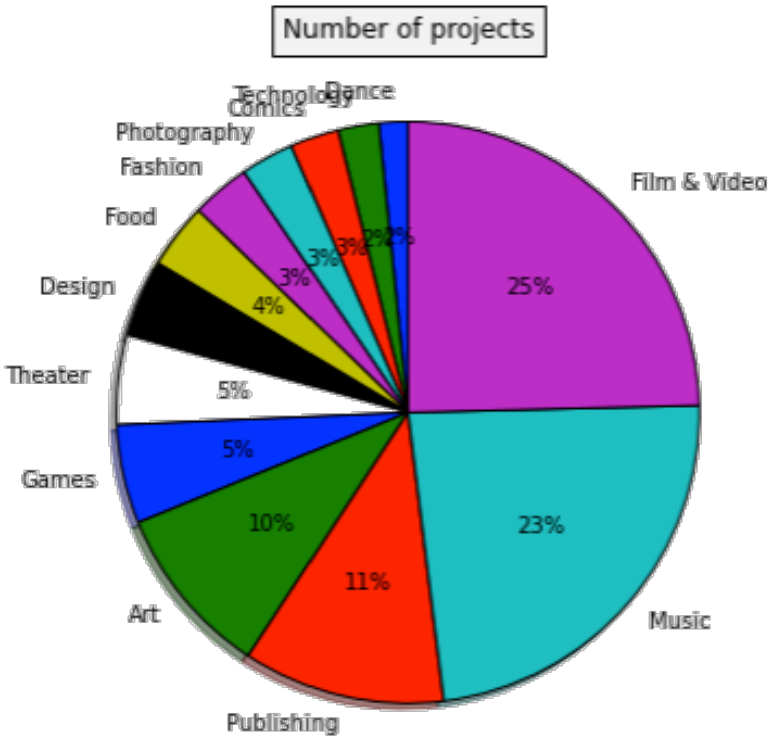
Number of projects per category

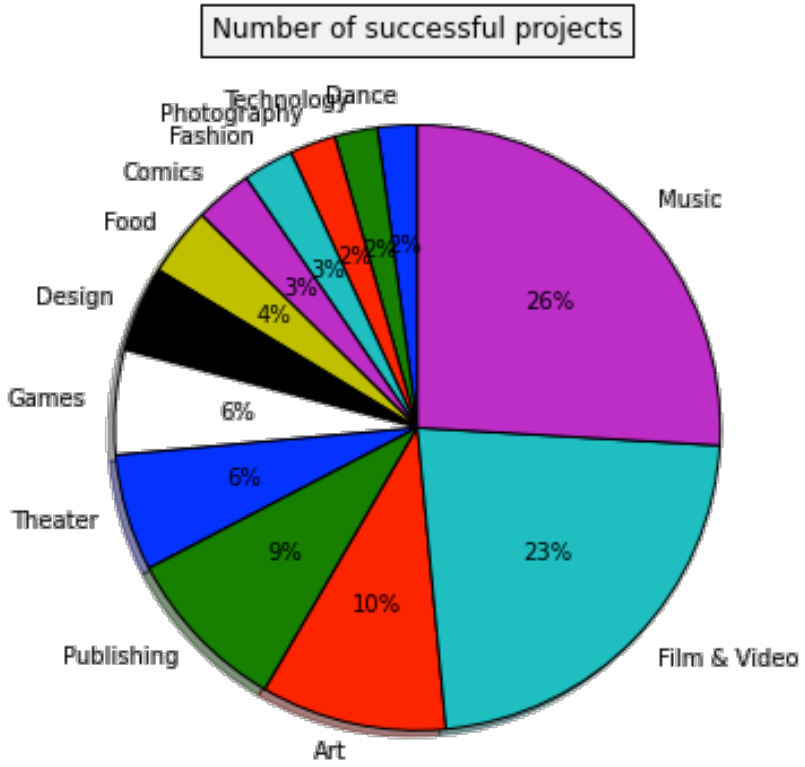


Analysis of sample of 60,000 Kickstarter projects
Source: RN analysis on data from kickstarter.com

… and the distribution is generally the same for successful projects

Number of projects per category

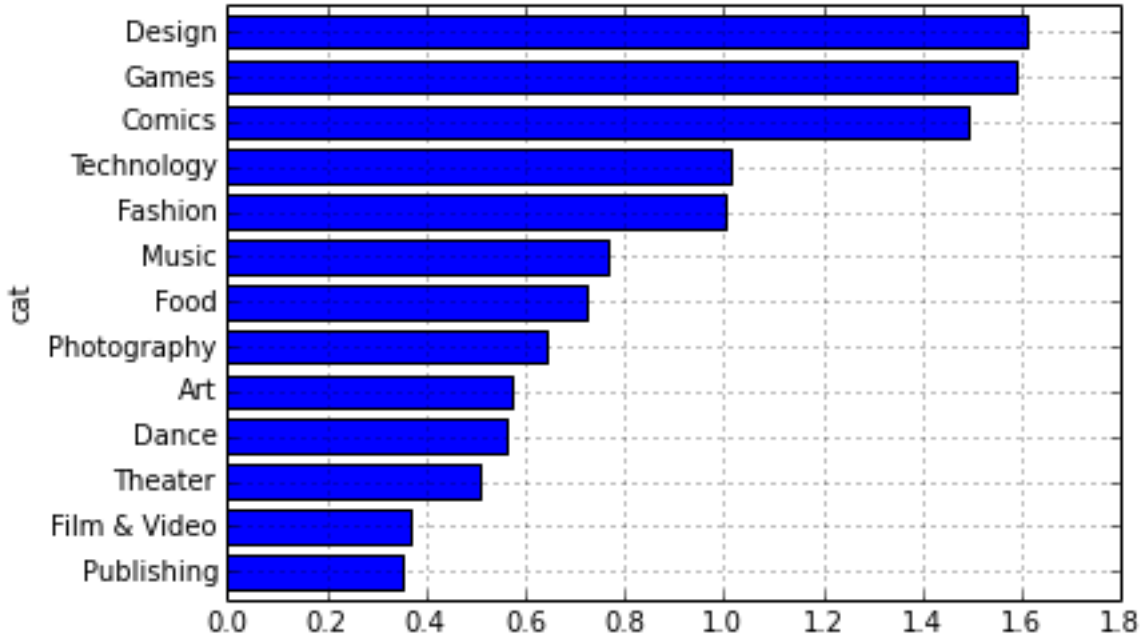




Analysis of sample of 60,000 Kickstarter projects
Source: RN analysis on data from kickstarter.com

Projects in Design, Games and Comics tend to raise 50-60% more than their goal…

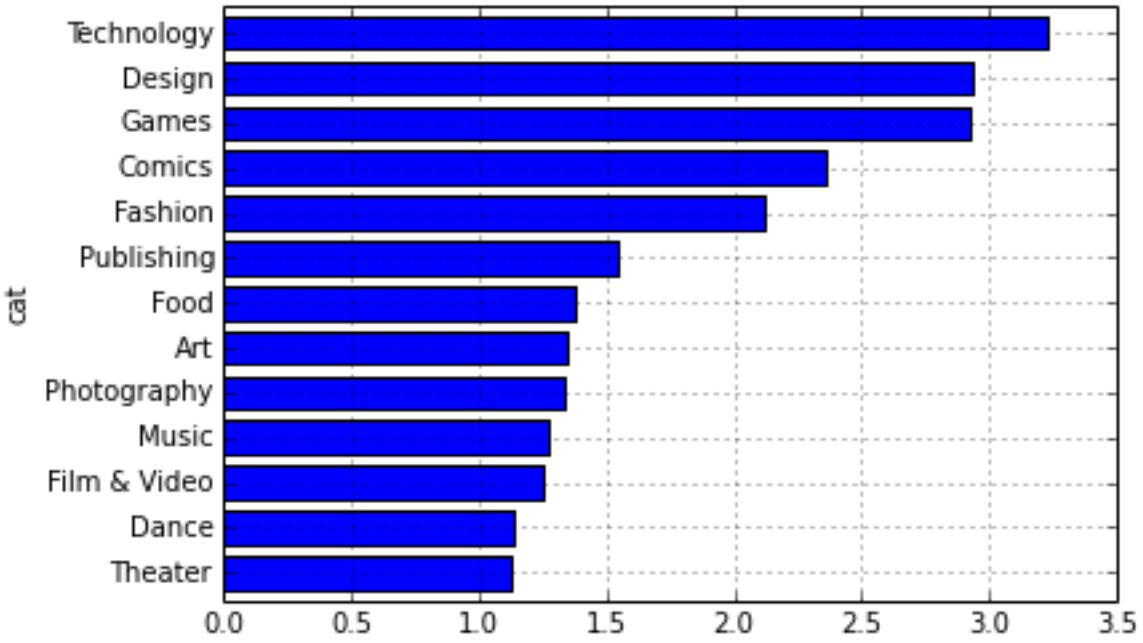
Funding rate per category (pledged/goal, including both successful and failed projects)



Analysis of sample of 60,000 Kickstarter projects
Source: RN analysis on data from kickstarter.com

… but successful Technology projects raise more than *three* times their goal

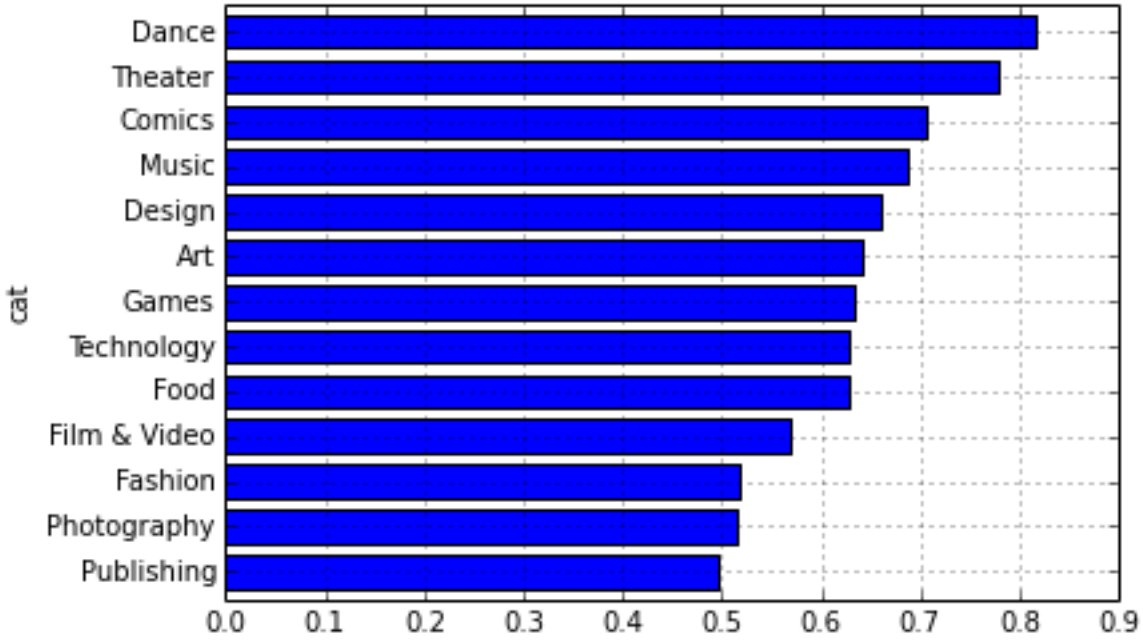
Funding rate of successful projects per category



Analysis of sample of 60,000 Kickstarter projects
Source: RN analysis on data from kickstarter.com

Projects in Dance and Theater are 4 out of 5 times successful, while about half of all projects in Fashion, Photography and Publishing fail

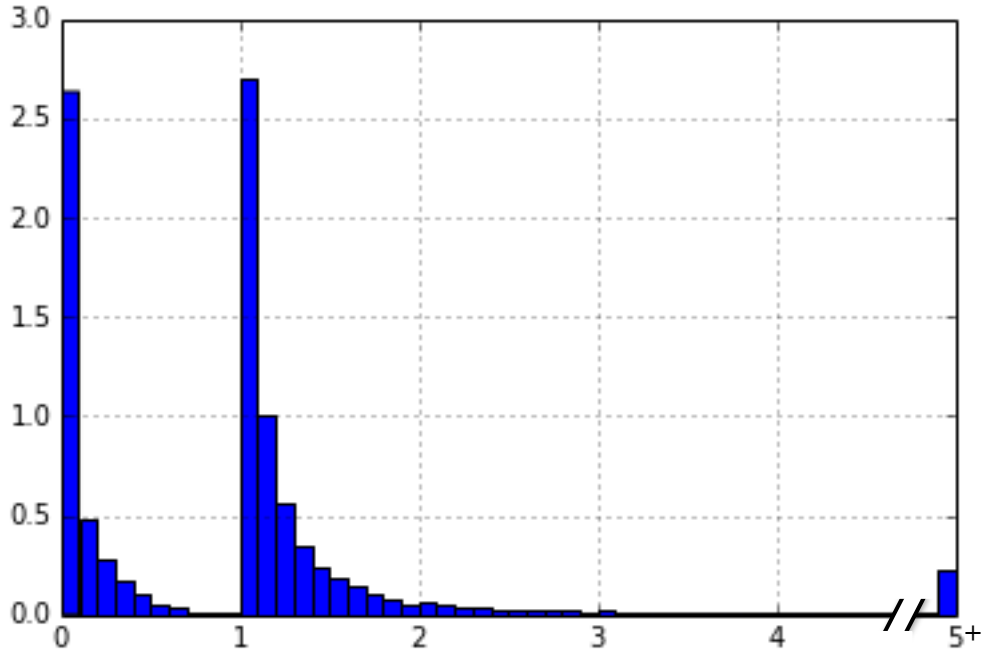
Success rate per category



Analysis of sample of 60,000 Kickstarter projects
Source: RN analysis on data from kickstarter.com

Most projects receive either next to nothing, or just enough.
Barely any project raises just under, or more than twice, their goal

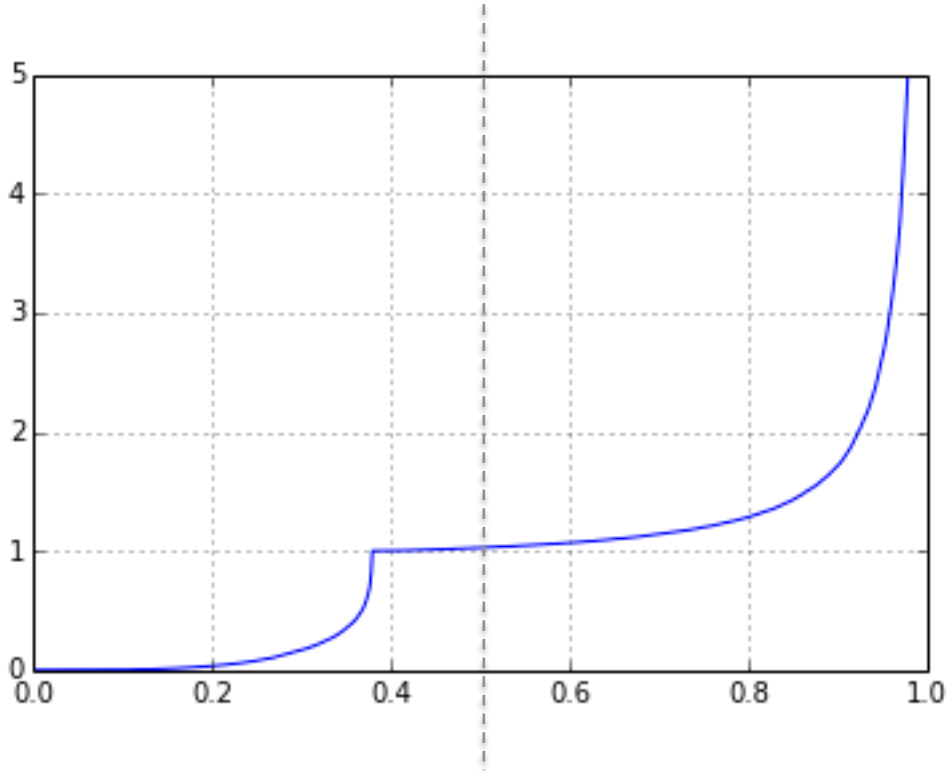
Distribution of funding rate (total area of bars = 1)



Analysis of sample of 60,000 Kickstarter projects
Source: RN analysis on data from kickstarter.com

Less than 40% of projects fail, 20% raise just enough, and less than 10% raise more than twice their goal

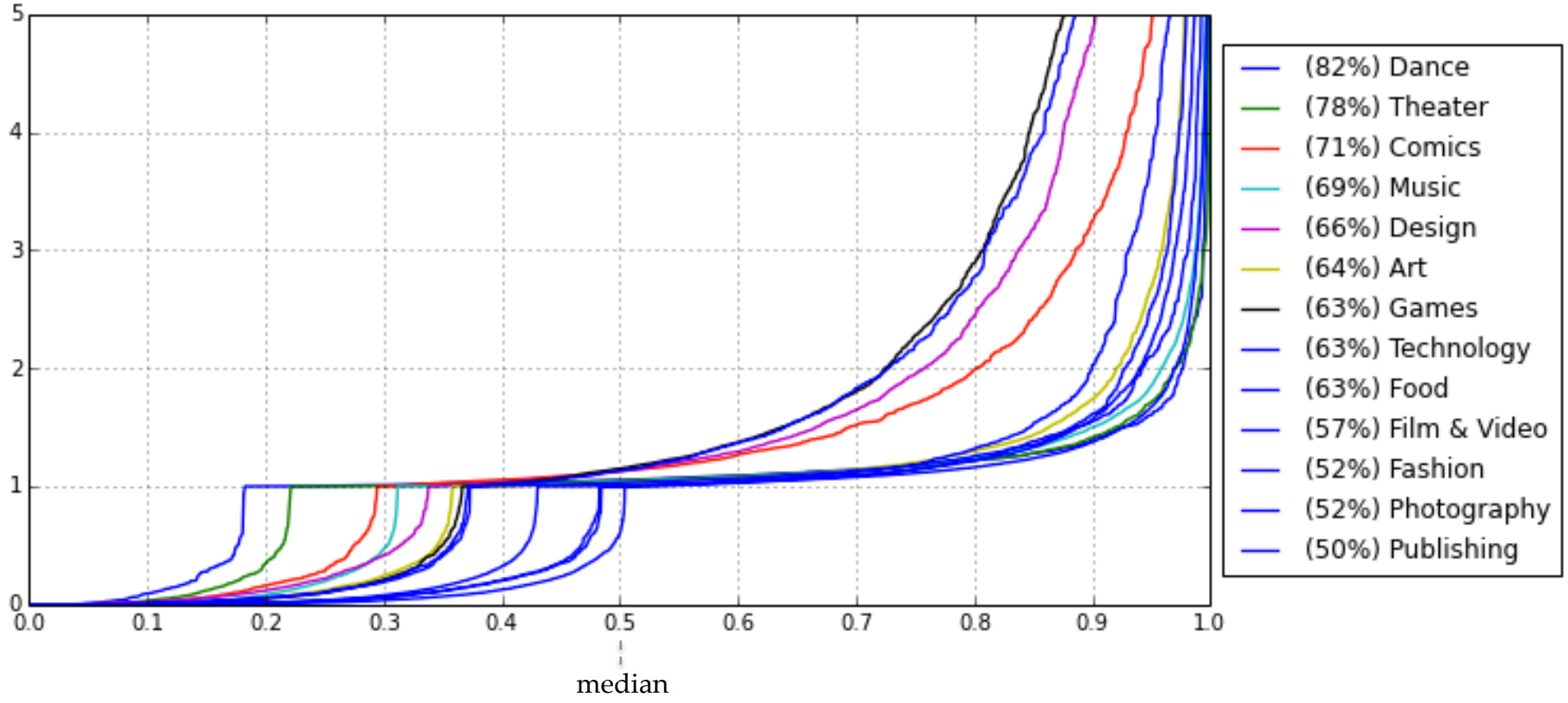
Distribution of funding rate (0.5 = median)



Analysis of sample of 60,000 Kickstarter projects
Source: RN analysis on data from kickstarter.com

Distribution of funding rate varies amongst categories

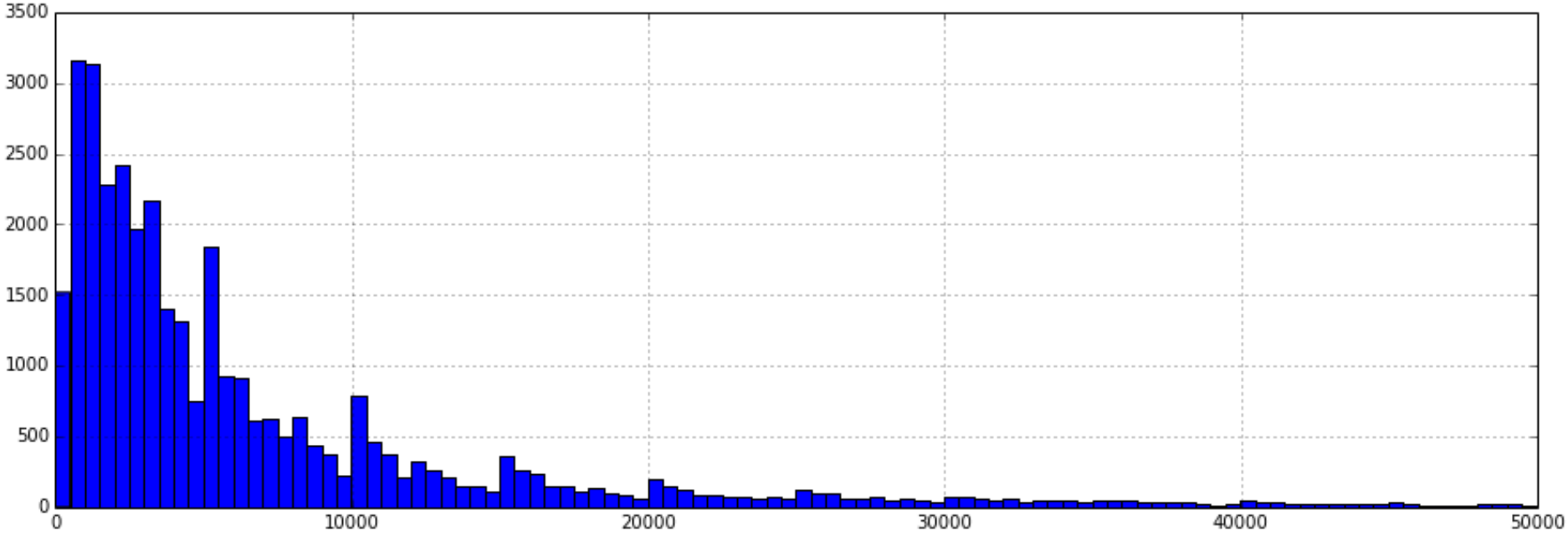
Distribution of funding rate, per category



Analysis of sample of 60,000 Kickstarter projects
Source: RN analysis on data from kickstarter.com

Most successful projects raise around $5,000 or less

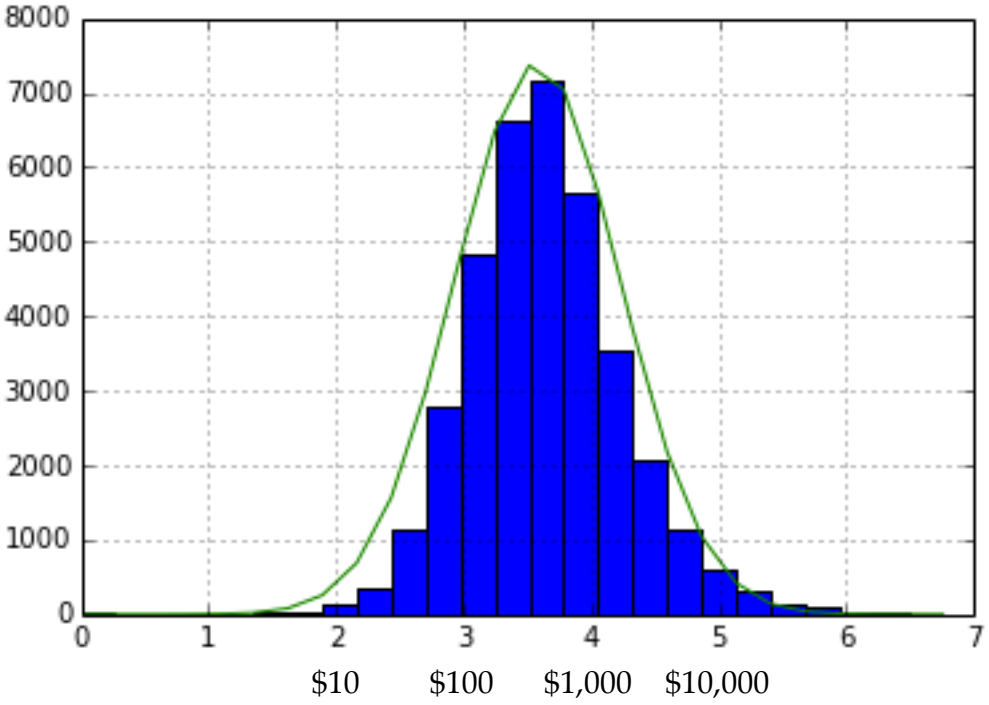
Distribution of pledges per successful project (one bar = $500)



Analysis of sample of 60,000 Kickstarter projects
Source: RN analysis on data from kickstarter.com

The successful pledges seem to be normally distributed on a logarithmic scale

Distribution of pledges per successful project (logarithmic scale, one bar = 0.2)



Analysis of sample of 60,000 Kickstarter projects
Source: RN analysis on data from kickstarter.com

Agenda

Descriptive Statistics

Machine Learning

First analyses were done with 15 features, of which many described the style of the text (rather than the text itself)

Selected features for predictions

Feature	Definition
title_nchars	# characters in the title
title_nwords	# words in the title
title_capsratio	% upper case (of any case) in title
desc_nchars	# characters in the description (tagline)
desc_nwords	# words in the description (tagline)
desc_capsratio	% upper case (of any case) in description
text_nchars	# characters in the body text
text_nwords	# words in the text
text_capsratio	% upper case (of any case) in the text
text_pars	# paragraphs in the text (“<p>”)
text_links	# links in the text (“<a ”)
text_imgs	# images in the text (“<img ”)
text_complexity	# characters per word, on average, in text
text_structured	# paragraphs per word, on average, in text
goal_log	logarithm of the goal (in any currency)

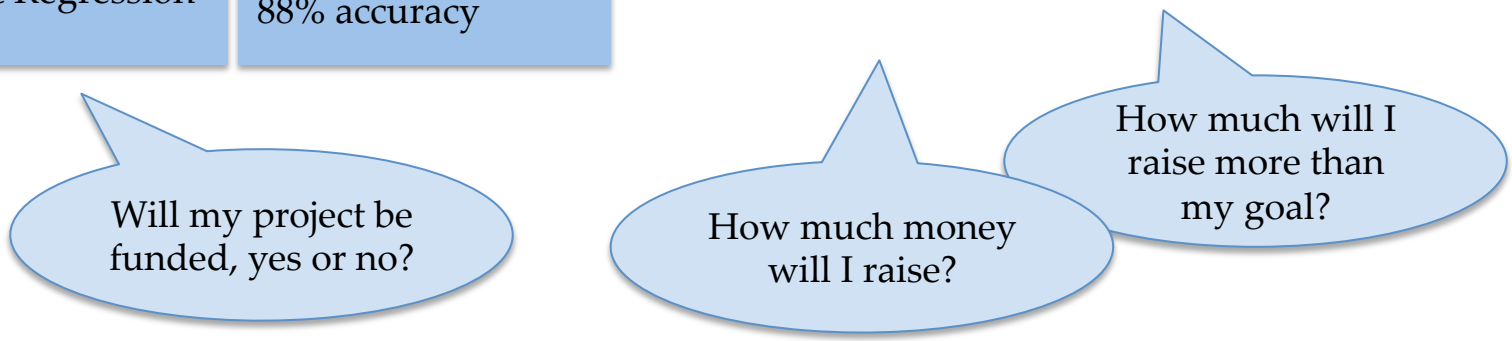
Scaled to a 0..1 range for equal comparison

Predicting success seems to work very well, while predicting pledges or funding ratio seems hopeless

Selected features for predictions

Predicting success seems to work very well…	
Model	**Cross validation**
Random Forest *(GradientBoostingClassifier)*	98% roc_auc*
Logistic Regression	89% roc_auc 88% accuracy

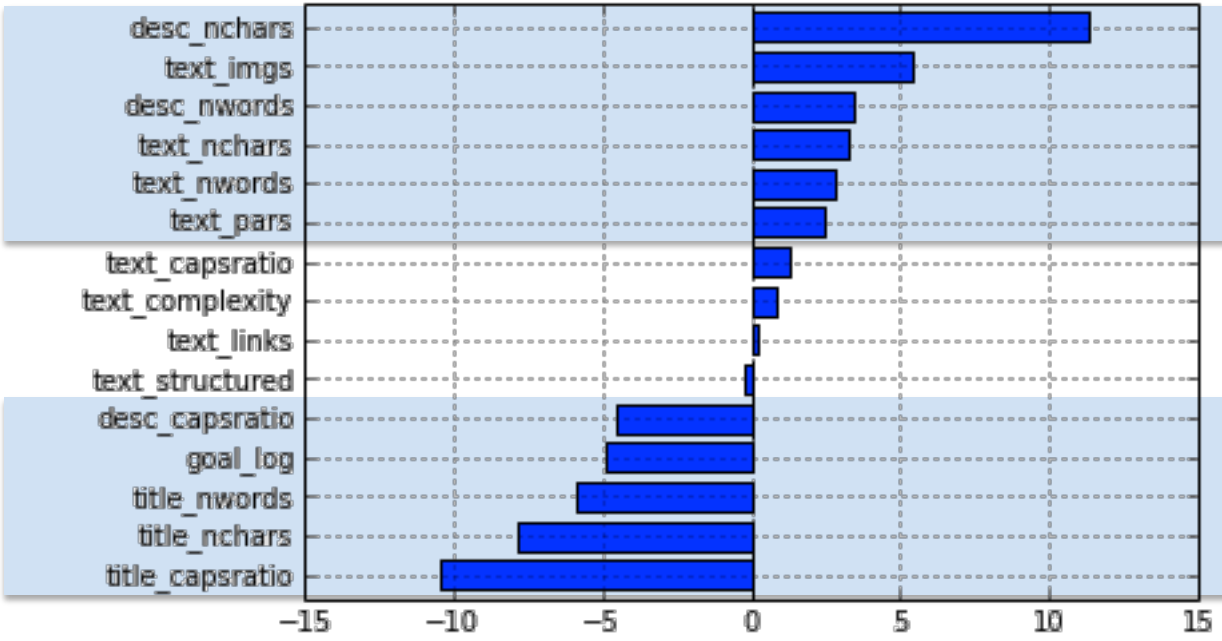
…while predicting pledges or funding ratio seems hopeless	
Model	**Cross validation**
Linear Regression	Very bad (R^2 or MSE < 0)



* Receiver operating characteristic, area under curve
Source: RN analysis on data from kickstarter.com

Successful projects have a long tagline and text with many images, while failing projects use many caps, have a long title and set a high goal

Impact of scaled features on success of project (logistic regression)

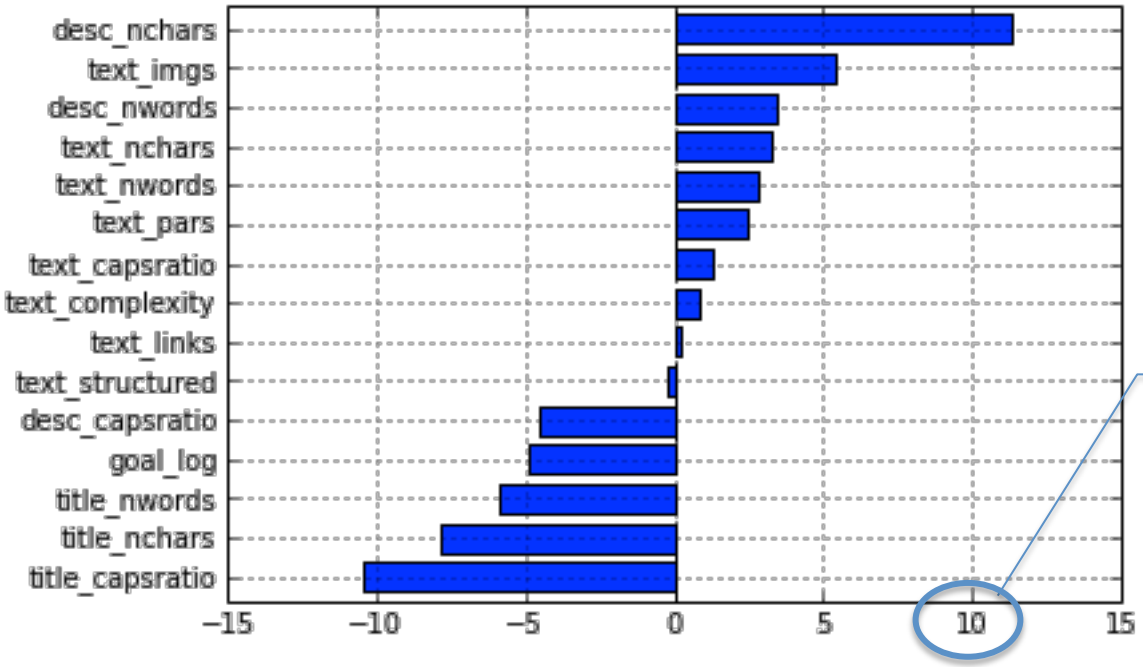


Do's
- a long text & tagline
- many images

Don'ts
- caps in title & tagline
- a high goal
- a long title

Analysis of sample of 60,000 Kickstarter projects
Source: RN analysis on data from kickstarter.com

Impact of scaled features on success of project (logistic regression)



A coefficient of 10 means that the project with the most characters (images, etc.) has

e^{10} = 22,000

as much chance of being successful than the project with the least characters

Note that features are scaled to [0..1], so these coefficients read as best versus worst. Maybe better would have been to scale to [-1..1] with median = 0.

Analysis of sample of 60,000 Kickstarter projects
Source: RN analysis on data from kickstarter.com

Regression on words counts in title, tagline or text *does* work, but not as well as the aforementioned metadata…

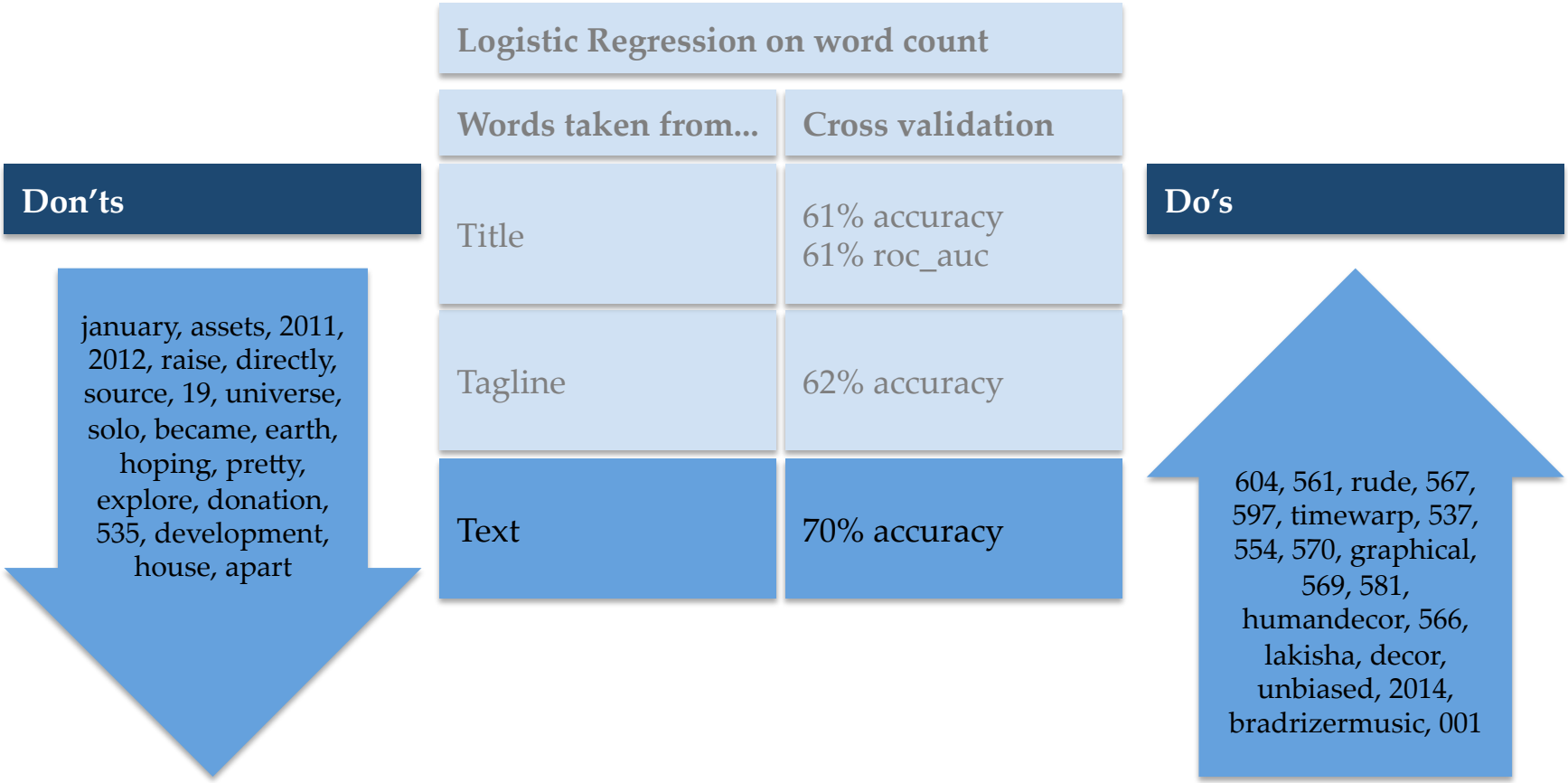
Logistic Regression on word count	
Words taken from...	**Cross validation**
Title	61% accuracy 61% roc_auc
Tagline	62% accuracy
Text	70% accuracy

Earlier roc_auc scores:
random forest 98%
logistic regression 89%

Analysis of sample of 60,000 Kickstarter projects
Source: RN analysis on data from kickstarter.com

… but the words that matter seem very arbitrary

Words from body text with lowest and highest coefficients in logistic regression



Analysis of sample of 60,000 Kickstarter projects
Source: RN analysis on data from kickstarter.com

Most important question is, though,
how can I adjust my project to improve my chances?

Example

I'd like to raise something between
$3,000 and 10,000 for my music project.
What goal should I set?

To compute your odds, train your model on a subset of similar projects

Example

I'd like to raise something between $3,000 and 10,000 for my music project. What goal should I set?

Steps

Select projects with
- Goal between $3 and 10K
- Category music

Train model on these projects

Take current project
- vary its goal between range
- predict probability of success

The subset-model has similar scores to the model that was trained on the entire dataset

Example

I'd like to raise something between $3,000 and 10,000 for my music project. What goal should I set?

Steps

Select projects with
- Goal between $3 and 10K
- Category music

Train model on these projects

Take current project
- vary its goal between range
- predict probability of success

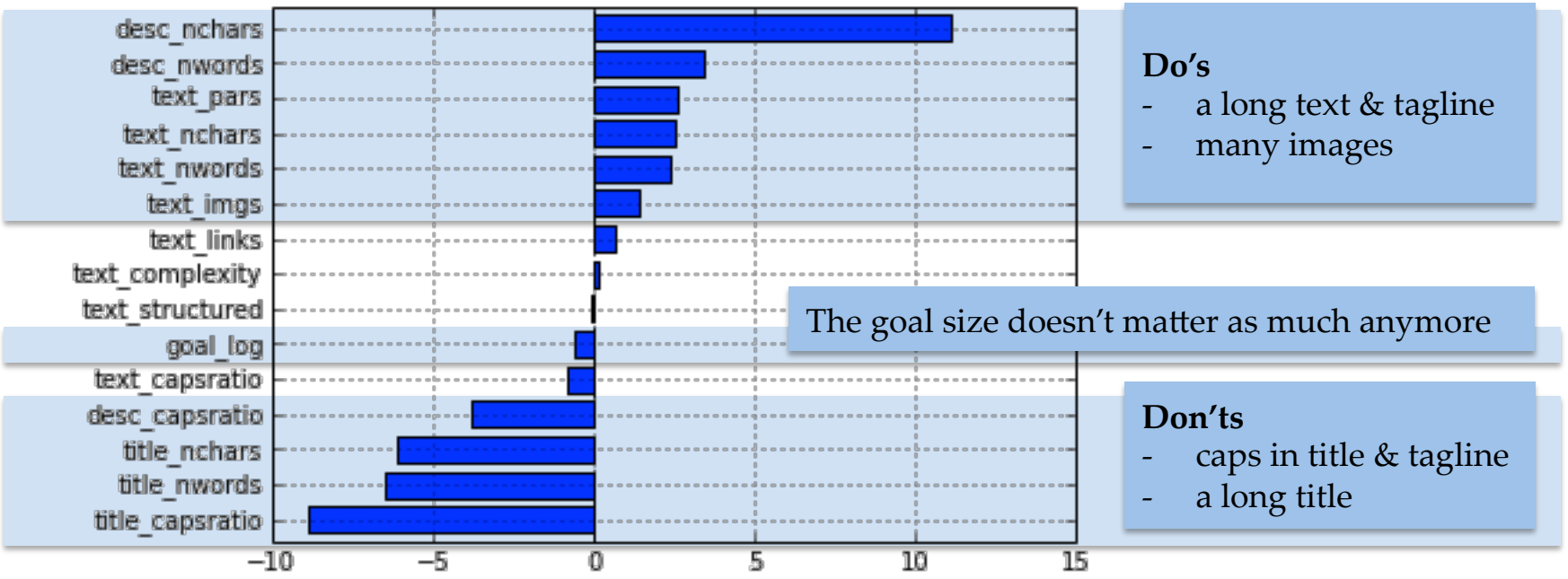
Results

Logistic regression has a cross validation score of 89% (roc_auc) or 87% (accuracy)

Analysis of sample of 60,000 Kickstarter projects
Source: RN analysis on data from kickstarter.com

Subset of similar projects give roughly same results as whole dataset, but the size of the goal doesn't have as much of an impact aymore

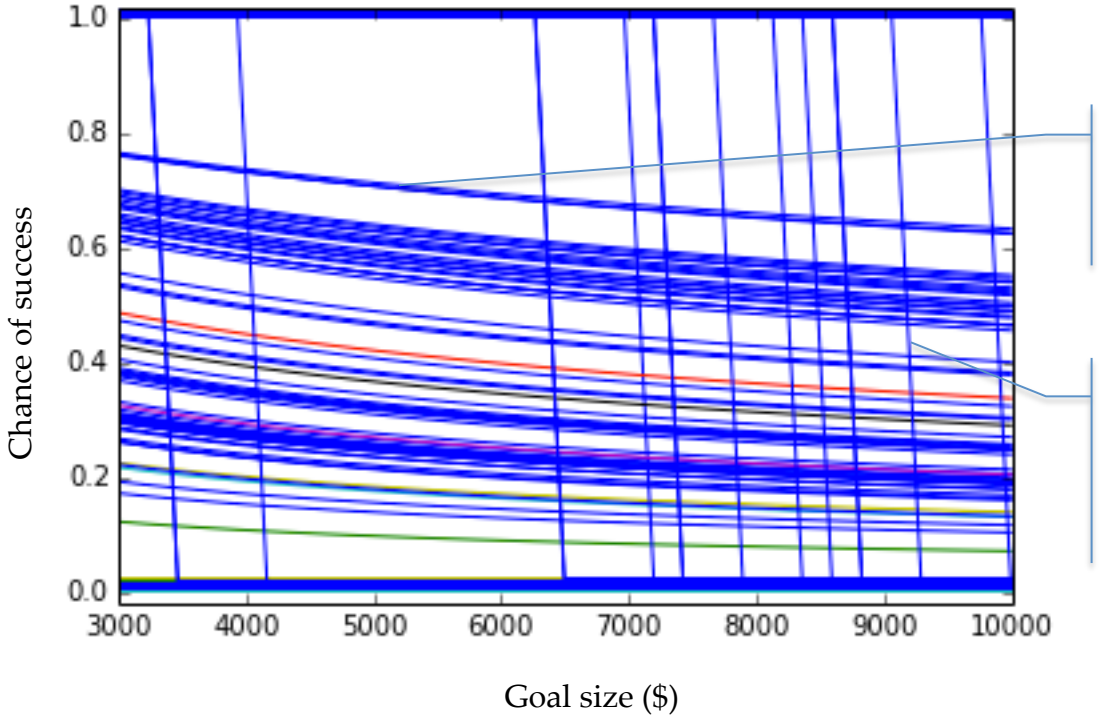
Impact of scaled features on success of project (logistic regression)



Analysis of sample of 60,000 Kickstarter projects
Source: RN analysis on data from kickstarter.com

Duplicating your ideal project with different values for your goal, tells you your odds for each goal

Success rate per category



Curves

Each curve represents one project: as its goal increases, its chance of success decreases

Lines

Each line is this chance rounded to 0 or 1, so the vertical strokes indicate the 50%-threshold

Analysis of sample of 60,000 Kickstarter projects
Source: RN analysis on data from kickstarter.com

What's next?

Open problems

- Add more features: location, rewards (text, prices, quantity, distribution), etc..
- Predict number of backers, average pledge per backer
- Vary with features and show plots how adjusting the feature will impact your chances
- Develop website that helps people launching their Kickstarter project