Research Overview

- Questions about products, especially new ones
- Why products succeed, which features are important
- This product/feature thread connects my projects:
  1. Attributes used for consideration set formation
  2. Hardware and app complementarity
  3. Serial creators on Kickstarter
  4. YouTube videos as "products"
  5. Improving ideation
  6. New work focusing on design and aesthetics

Is Creativity Purely Random? Testing Alternate Algorithms for Idea Screening in Crowdsourcing Contests

J. Jason Bell, Christian Pescher, Gerard Tellis

Problem

- Crowdsourcing contests generate many ideas
- Currently screening is done by experts, but this is costly
- Can algorithms help?
- Depends whether there are stable patterns in creative ideas

The Value of Expert Judgements

- Recent evidence suggests expert judgements are useful for cases where:
  - Depth of knowledge helps narrow scope to more feasible ideas
  - Very little information exists (a pitch or other high level description)
  - Far in advance of deployment (speculating vs. reacting)
- Community or consumer judgements are more accurate for the opposite cases

Literature

- Toubia & Netzer (2016) metrics from word colocation networks, derived from a theory of balance between novelty vs. relevance
- Totally new metrics, hand-crafted for the task of evaluating ideas
- Method requires a "reference corpus," which is a group of documents about the contest topic
- Generates many variables to use as predictors: each pair of words has several metrics, these are aggregated up to idea level

Literature

- Stephen, Zubscek and Goldenberg (2016) clustering coefficient impact on idea quality
- How many of your friends are friends with each other?
- Clustering coefficient negatively related to diversity of information
- Lower CC $\rightarrow$ more diverse information/inspiration $\rightarrow$ better ideas

Literature

- Berger and Packard (2018) topic atypicality positively related to popularity
- Within a particular genre, songs whose lyrics are atypical may stand out more
- Topic models used to determine content, style matching to measure similarity
- Conditional on genre, more atypical lyrics related positively to popularity

Our Setting and Goal

- Data from 10 idea contests. Contest process:
  1. Contest brief posted
  2. Community can view ideas, comment, and give ratings
  3. Experts select shortlist
- Can we reduce work for the experts with a model?
- Standard given by Hyve:
  1. Eliminate 25% of the worst ideas
  2. Sacrifice no more than 15% of shortlisted ideas

Contributions

- Real world data, out-of-sample, different DV
- Extend all three methods tested:
  - Toubia & Netzer: new text corpora used as reference corpora
  - Stephen et al: added Burt's measure of constraints
  - Berger & Packard: new measure of topic atypicality
- Prediction accuracy good enough for use, better than human benchmarks
- Find very predictors across wide range of contests

Methods

- Toubia & Netzer's variables with different source texts:
  - Google (Toubia & Netzer, 2016)
  - Wikipedia (new)
  - Patents (new)
  - Own-contest ideas (new)
- Clustering coefficient (Stephen et al, 2016) and Burt's measure of constraints (new)
- Topic atypicality using LDA (Berger and Packard, 2018) and topic overlap (new)

Lasso Logistic Regression

- LHS Variable:

$$y = \begin{cases} 1 & \text{if the idea was shortlisted by experts} \\ 0 & \text{otherwise} \end{cases}$$

- RHS variables:
  1. Vars from previous lit + our extended versions
  2. Each sample estimated wiith and without community ratings

Lasso as a Variable Selection Tool

- Lasso is a ‘regularized’ linear model.
  - Regularized = includes an explicit bias
  - For Lasso, bias is toward parsimony
- Lasso sets some variable coefficients to zero
- Remaining predictors often quite robust

Lasso in a Logistic Regression

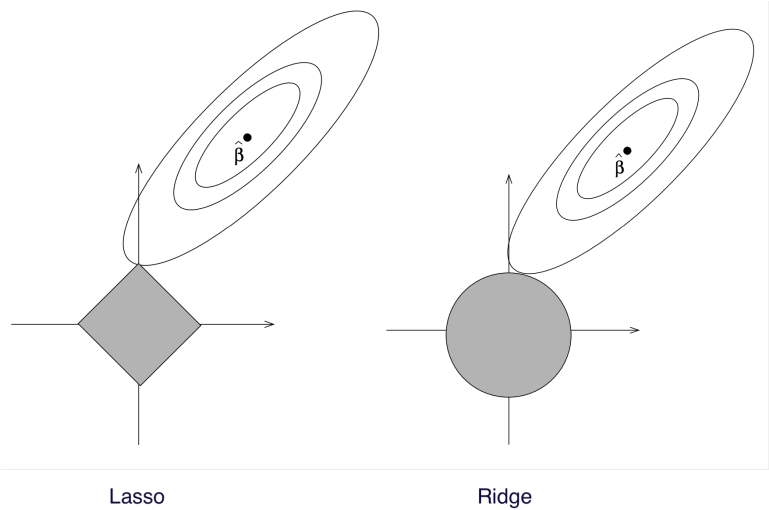
- Lasso objective function for continuous y_i:

$$(\hat{\beta}_0, \hat{\boldsymbol{\beta}}) = \arg\min_{\boldsymbol{\beta}} SS(\beta_0, \boldsymbol{\beta}) + \lambda||\boldsymbol{\beta}||_1$$

$$SS(\beta_0, \boldsymbol{\beta}) = \sum_{i=1}^{N}(y_i - \beta_0 + \boldsymbol{x}_i'\boldsymbol{\beta})^2$$

- Least squares with a "budget"

Geometry of Lasso Compared to Ridge



Lasso in a Logistic Regression

- Small modification for binary LHS.
- Objective is penalized MLE, instead of penalized least squares:

$$(\hat{\beta}_0, \hat{\boldsymbol{\beta}}) = \arg\max_{\boldsymbol{\beta}} \ell(\beta_0, \boldsymbol{\beta}) + \lambda ||\boldsymbol{\beta}||_1$$

$$\ell(\beta_0, \boldsymbol{\beta}) = \frac{1}{N} \sum_{i=1}^{N} \left[y_i(\beta_0 + \boldsymbol{x}_i' \boldsymbol{\beta}) + \log \left(1 + e^{\beta_0 + \boldsymbol{x}_i' \boldsymbol{\beta}} \right) \right]$$

Why Lasso?

- Lasso is good at finding robust predictors
- We face small datasets, wide range of domains (from plastic extrusion to customer service apps)
- Variables all have previously established meaning, we care about coefficient values
- An ensemble of RF, XGboost, Lasso seems to work marginally better, wasn't worth complexity and loss of intepretability

Overall Sample

Contest	N. Ideas	N. Users
DHL	131	79
FR	495	118
Fujitsu	218	76
Ideabird	595	235
Lufthansa1	238	100
Lufthansa2	145	80
Mastercard	143	48
Reifenhaeuser	55	17
Vodafone	161	52
Zeiss	116	66
Total	2297	871

Estimation Samples

- One in-sample model with all 10 contests
- Out-of-sample models for each contest:
    - Nine contests used for training
    - Remaining contest used to compute out-of-sample accuracy

Class Imbalance

- Only around 6% of the observations are shortlisted
- Model doesn’t get to see many shortlisted ideas
- Expert ratings should help but don’t
  - Many idiosyncrasies across contests mapping from ratings to shortlist
  - Experts can shortlist ideas unilaterally, regardless of ratings
  - Expert ratings are quite sparse
- We use SMOTE (synthetic minority oversampling technique) to partially mitigate class imbalance.

Variables Used by Lasso

Variable	Coefficient
Node Freq. Coef. of Variation (Wiki.)	-0.35
Community Ratings	0.34
Clustering Coefficient	-0.25
Constraints	-0.22
Topic Overlap Atypicality	-0.15
Minimum Node Frequency (Own)	0.13
Jaccard Index Coef. of Variation (Google)	0.12
Average Jaccard Index (Patents)	0.10
Jaccard Index Coef. of Variation (Own)	0.07
Average Node Freq. (Own)	0.02
Kolmogorov-Smirnov Metric (Patents)	0.02

Bold items novel to this work

Unpacking Coef. of Variation (Wikipedia)

- To get the Node Frequency Coef. of Variation (Wikipedia) for an idea:
  1. For each word, count occurences in the Wikipedia reference corpus
  2. Make a vector of the counts (length = number of words in idea)
  3. Take coef. of variation of that vector

$$CoV(\boldsymbol{v}) = SD(\boldsymbol{v})/Mean(\boldsymbol{v})$$

- Ideas with high CoV are less likely to be shortlisted.

What Leads to High CoV (Wikipedia)

1. Irrelevance (small denominator)
2. Scattered (array of frequent and infrequent words as measured by occurences in Wikipedia)

How to Have Bad Ideas

- Miss the trends
- Talk about fringe stuff
- Be complicated or scattered
- Propose obvious, common things

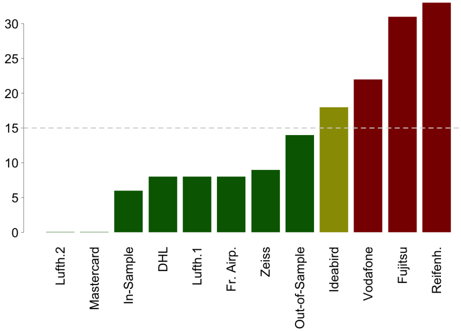
Evaluation Metrics

- Usually we’d use something like an F1 score: $\frac{2pr}{p+r}$
- Here, manager cares about precision conditional on removing 25% of ideas: we call this "the sacrifice rate".
- We want to hit a sacrifice rate of 15% or lower.

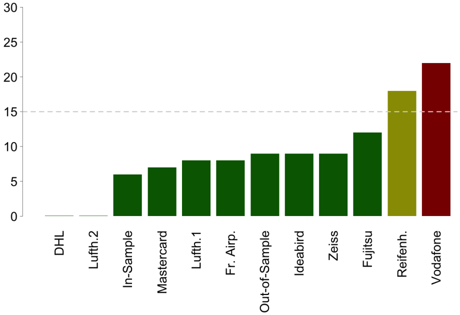
Is 15% Easy?

- We thought it would be, at first
- It turns out to be quite difficult for reasons discussed:
    1. Really wide range of domains
    2. Small, incomplete data
    3. Class imbalance
    4. DGP unstable
- Community ratings can provide a "human benchmark" to tell us how hard this really is

Sacrifice Rates by Sample

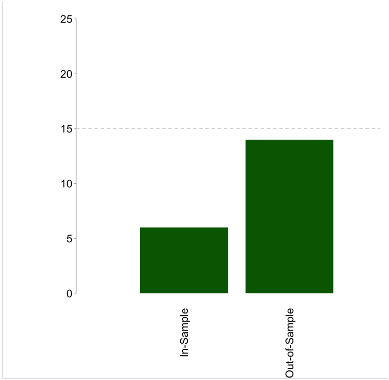


Without Community Ratings

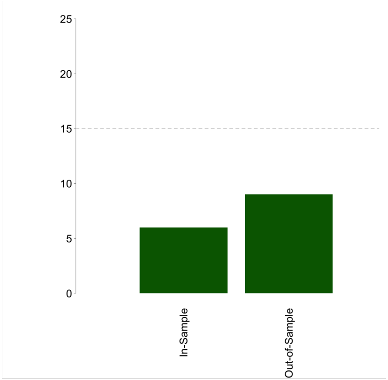


With Community Ratings

Performance on Aggregated Samples



Without Community Ratings



With Community Ratings

Benchmarks: Random and Community

- Best model meets threshold in 8/10 contests
- Odds of getting 8/10 by chance $<$ 1 in 10,000
- With only community ratings: 5/10
- With everything but community ratings: 6/10
- Algorithms together do better than collective community judgement (and they are really easy to run!)

Contributions

- Real world data, out-of-sample, different DV
- Extend all three methods tested:
    - Toubia & Netzer: new text corpora used as reference corpora
    - Stephen et al: added Burt's measure of constraints
    - Berger & Packard: new measure of topic atypicality
- Prediction accuracy good enough for use, better than human benchmarks
- Find very predictors across wide range of contests

What Next?

- Received 9 more contest datasets
- Paper will improve with more data, delay in submission is minor
- Submission target is *Marketing Science* by end of November

Kickstarter Paper Summary

- **Empirically** show very successful Kickstarter campaigns lead to higher goals next time
- Big success → higher next goal → more likely to fail
- Replicate effect in the **lab**, show evidence of anchoring as cause
- New, realistic anchor removes the effect