

Predicting Loan Fundraising Success

5/31/2016
Emelyn Chew



Why?

1. Kiva Zip is a P2P crowdfunding platform that makes 0% interest rate loans to small business owners
2. All-or-nothing model: loans expire if they do not reach fundraising target within 60 days
 - Borrowers do not get any \$ 😞
 - Lenders get money back in their Kiva account 😞
 - Operational cost 😞

Question: Can we predict if a borrower will successfully fundraise on the Kiva Zip platform given certain loan characteristics?


Getting the data

1. Scraped Kiva Zip Website using Beautiful Soup
2. Predictors:
 - Business name
 - City
 - State
 - Conversations with lenders
 - Lenders (new and invited)
 - Text: Personal Story, Business Description, Loan Purpose
 - Status
 - Loan Size
 - Industry
 - Years of experience
 - Social Media + Website
 - Trustee

Ginger

R-evolution Gardens
Nehalem, OR

[OVERVIEW](#) [ENDORSEMENT](#) [LENDERS 186](#) [CONVERSATIONS 16](#)



Personal Story

Ginger Edwards is the head organic farmer and founder of R-evolution Gardens. What began in 2008 as a market garden on a 1/2 acre of heavy wet clay and tree stumps has grown under her care to feed to a 55 member CSA, a weekly farmers market, year round farm stand and many coastal wholesale accounts from a lush 3 acres of diverse organic vegetables.

Ginger Edwards was born in Detroit, Michigan and moved to Oregon in 1998 after graduating with a degree in Environmental Science from Western Michigan University. She began her stay in Oregon as a forest defense activist fighting illegal logging sales on Mt. Hood. Eventually her passion for sustainable living led her to growing organic food and feeding her community. She has lived on the Oregon coast since 2001 and in that time has opened two brick and mortar retail stores, helped found the Manzanita Farmers Market of which she served as Board Chair


PAYING BACK

A loan of \$10,000 helps me to grow and sell more vegetables to my local community.


Disbursed May 2, 2016

Repayment Term:	36 monthly payments
Grace Period:	6 months
Listed:	Mar 22, 2016
Disbursed:	May 2, 2016
Paying Back:	May 2, 2016

About R-evolution Gardens

Industry:	Agriculture
Years in Operation:	More than 5 years
Website:	revolutiongardens.com
Badges:	

This loan is endorsed by

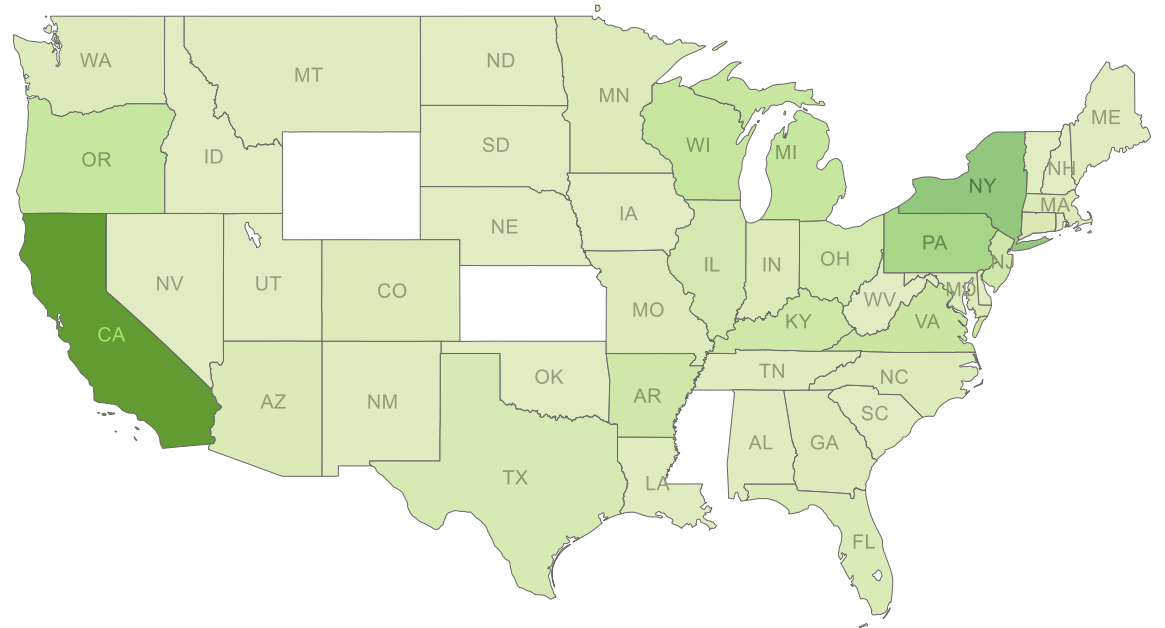


Rebecca Parker

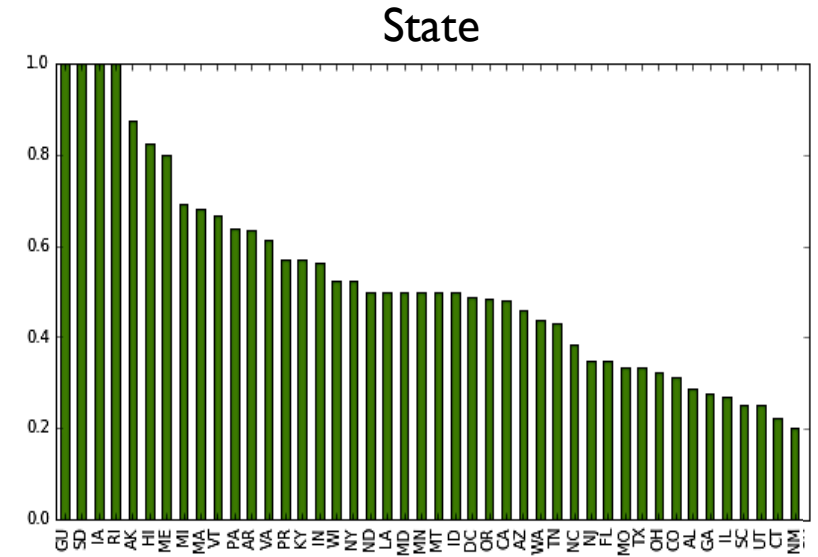
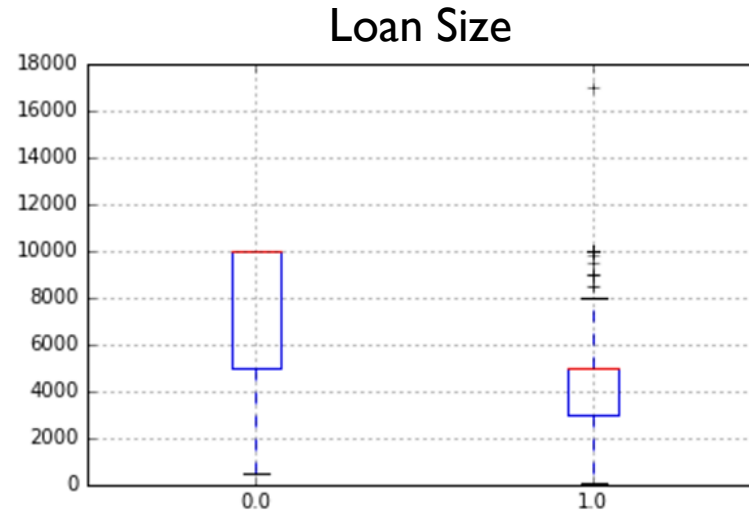
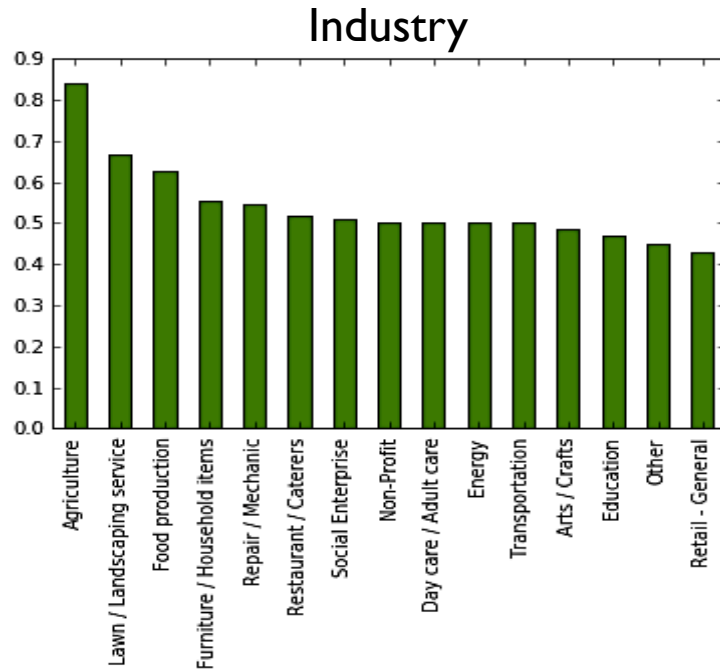
Trustee Type:	Individual
Tier:	Third Tier trustee
Location:	Portland, OR

What does the data look like?

1. +10,000 small business loans from 2011 to present
2. Filtered for U.S. based loans when scraping (~2500 loans)
3. 18 predictors
4. Response variable:
 - Days Fundraising: Difference between loan listed date and loan disbursed date
 - Eventually bucketed into 0 (failure), 1 (success) to become a binary classification problem



Initial useful features...



- Industry: Agriculture/Food production loans outperform
- Loan Size: Large loans have higher chance of failing
- States: Certain states have an advantage

Machine Learning Model #1: Logistic Regression

1. Trouble with linear regression: converted to binary variables to make the problem a classification problem
2. Null Accuracy Rate = 0.50
3. Logistic regression using features:
 - Loan size
 - Trustee (created dummy variables)
 - # Lenders invited
 - # Conversations with borrowers
 - Industry (created dummy variables)
 - Loan Term

Accuracy Rate = 0.62

NLP increased accuracy rate...

I. Added Natural Language Processing:

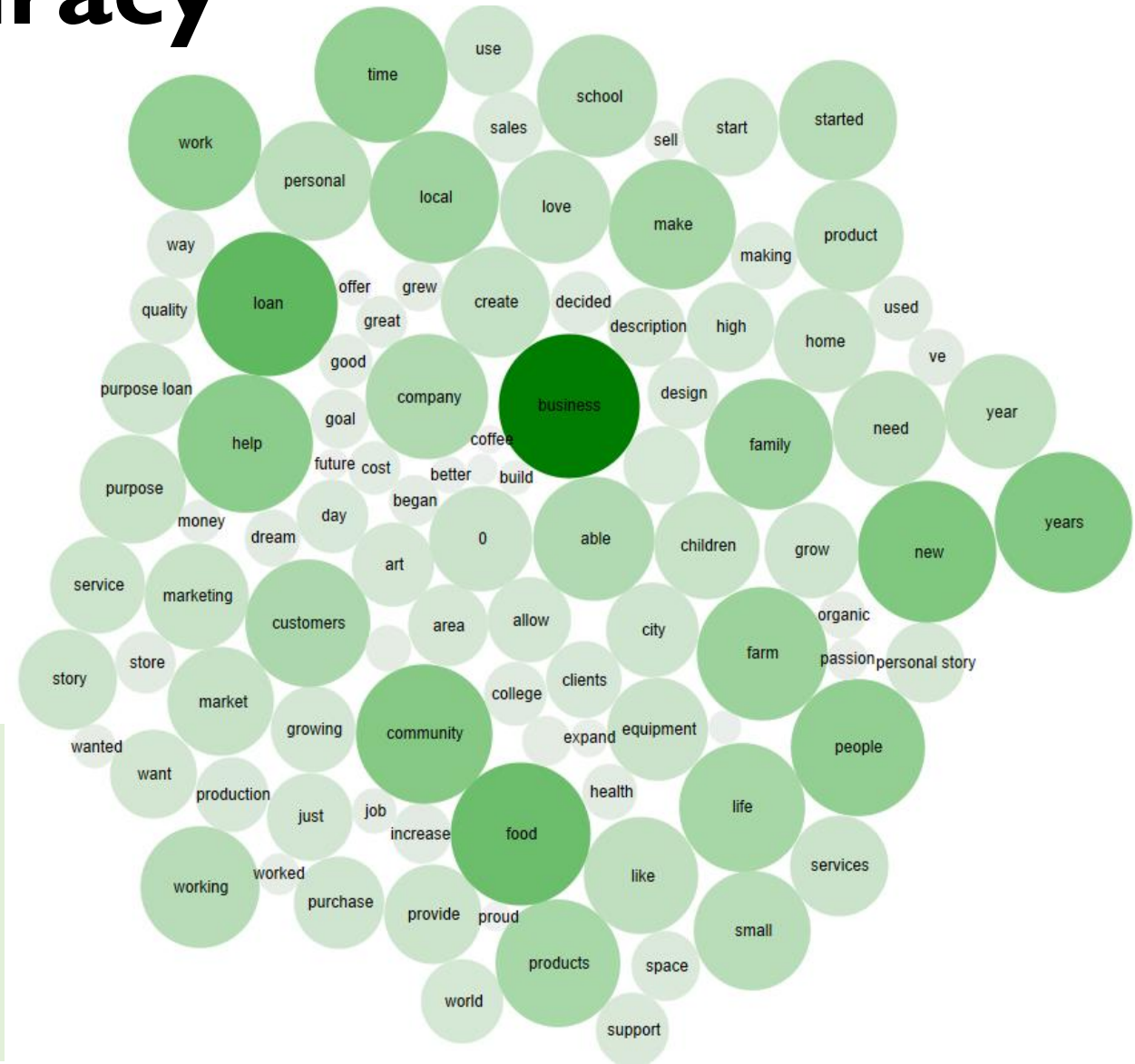
- TFIDF Vectorizer
- Ngram range (1,2)
- Max features = 1000
- Min df = 2
- Stop words

Accuracy Rate increased to 0.68

Sensitivity = 0.60

Specificity = 0.68

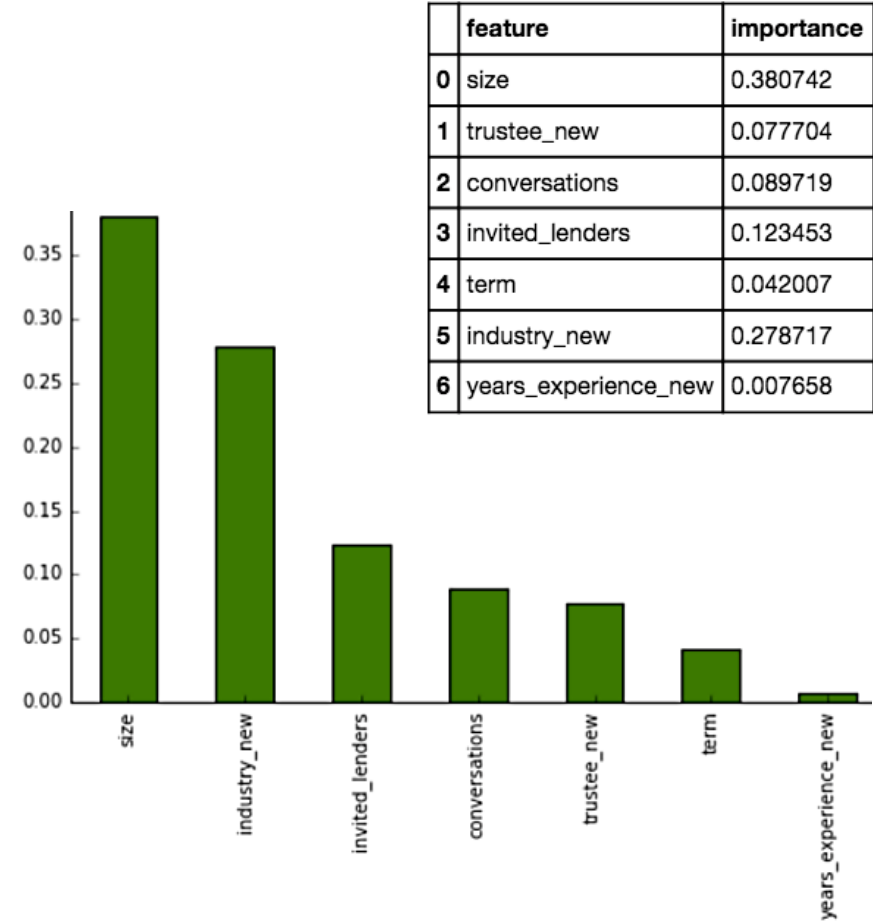
Model better at identifying *true negatives*!



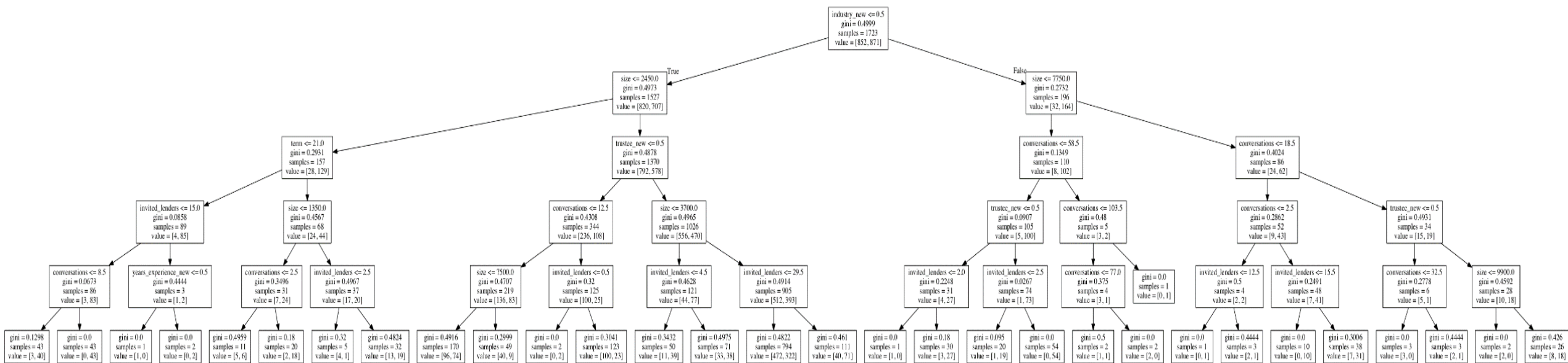
Machine Learning Model #2: Decision Tree

1. Tuning the regression tree:
 - Max Depth = 5
 - Cross validation
2. Gini index before split = 0.49
3. Size and industry most important features
 - Highest feature importance rates (reduce classification error rate by the most)
4. Trustee, term and years experience not that important

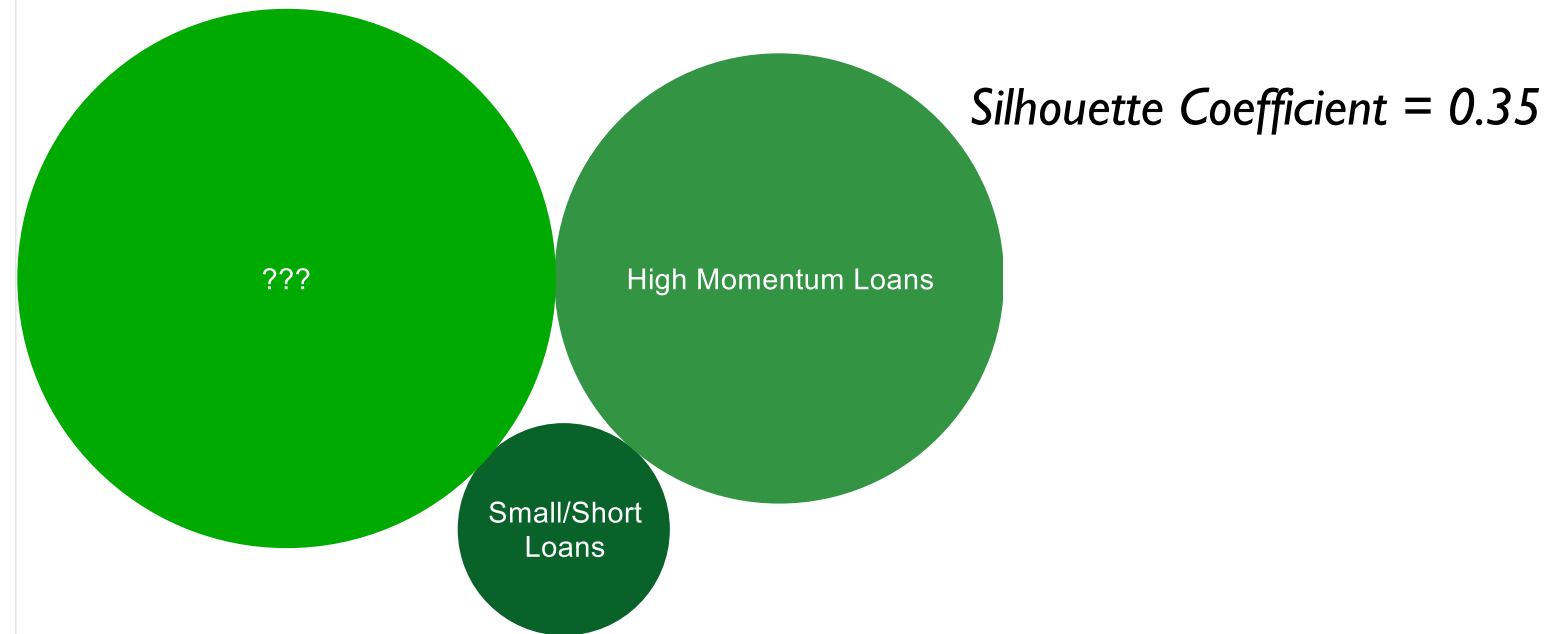
Accuracy = 0.74



Machine Learning Model #2: Decision Tree



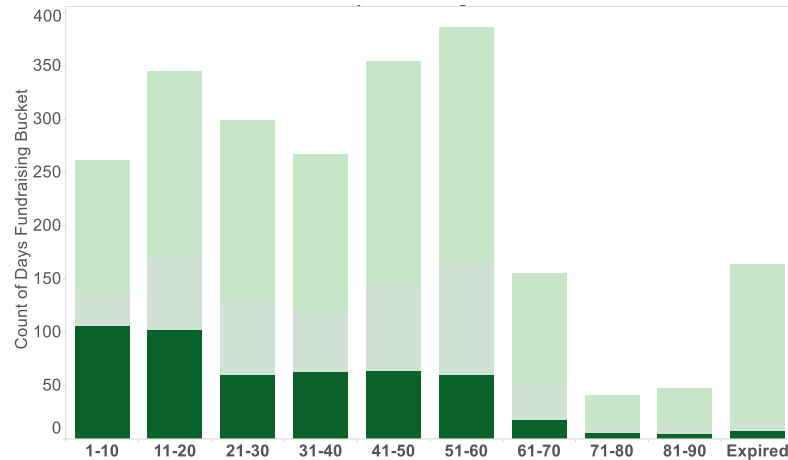
Machine Learning Model #3: K-Means Clustering



	size	state_new	trustee_new	conversations	invited_lenders	term	days fundraising	industry_new
cluster								
0	4986.653166	0.476436	0.770987	25.174521	14.968336	23.657585	45.128130	0.080265
1	9854.697286	0.638831	0.789144	19.956159	22.177453	34.739040	44.162839	0.273486
2	2148.159509	0.451943	0.660532	16.102249	9.259714	18.873211	32.057260	0.069530

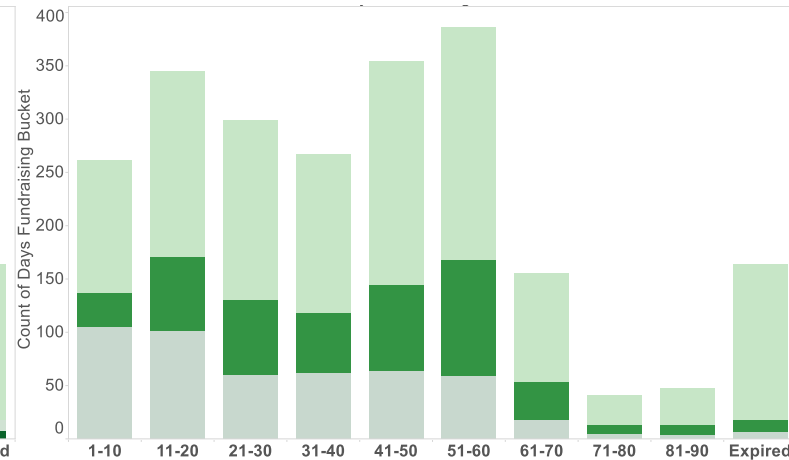
3 Distinct Clusters

Cluster 1: Days Fundraising



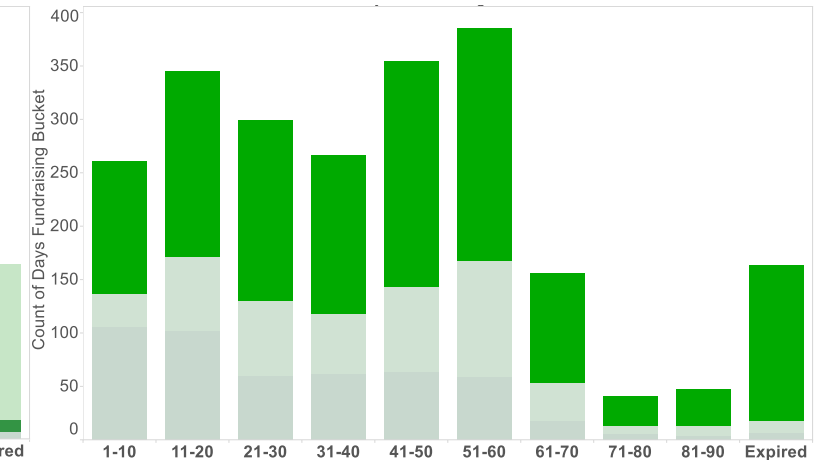
- Low Risk
 - Small loan size (~2k)
 - Short loan term (<20 months)
- Fundraise quickly, low effort

Cluster 2: Days Fundraising



- High risk
 - Large in size (~10k)
 - Long loan term (~36 months)
- High Momentum
 - CA/NY (Kiva HQ)
 - Trustee support
 - Many conversations & invited lenders
- End up succeeding at last minute

Cluster 3: Days Fundraising



- Average everything
 - Average size
 - Average term
 - Average borrower engagement
- Difficult to predict outcome

Conclusions

1. Fundraising success is hard to predict with high accuracy
2. Text matters!
3. Kiva staff:
 - Focus on loans in “Average” category – most likely to expire
4. Borrowers:
 - Choose realistic fundraising goals
 - Write good “stories”
 - Engage lenders in network
5. Given more time... more feature extraction:
 - Social media engagement – measure conversations about the loan on twitter/Facebook etc.
 - Time series data – measure if velocity of fundraising increases likelihood of success
 - Image processing – quality of photos, RGB, facial recognition