Thirteen ways of looking at a default

MAS.622J Final Project - Prosper.com

Charlie deTar
Coco Krumme
Ernesto Martinez-Villalpando
Matt Aldrich
review of questions

• What is the likelihood of a listing becoming a loan?

• What is the likelihood of a loan being paid on time?

• Which features best predict these outcomes?
why this is important

- Borrowers want to maximize likelihood of getting a loan, minimize interest rate
- Lenders want to invest in loans that maximize returns, minimize probability of default or late payment
- Prosper.com wants to maximize revenues by increasing loan conversion, decreasing default rate
- The research questions are deep: is peer-to-peer lending a viable model? How much do social factors matter? How do lenders make decisions? What models best capture loan dynamics? Can peer-to-peer be modeled with the same precision as bank loans? How can human classification aid machine learning algorithms? etc etc
roadmap

introduction desc. stats

feature selection

classification methods

graphical models

comparison of models

loan or no loan? default or pay?

greedy

PCA +NN

LDA

SVM

decision trees

bayes nets

other methods

HMM loan performance

mech turk

group performance
Prosper data tables

**Bid**
- Amount
- Minimum Rate
- Listing Status
- ...

**Category**
- Key
- Name
- Hierarchy
- ...

**Group**
- Member Key
- Group Rating
- City
- ...

**Listing**
- Amount Funded
- Amount Remaining
- Bid Count
- ...

**Loan**
- Credit Grace
- Borrow Rate
- Debt to Income Ratio
- ...

**Member**
- Key
- Friend Member Keys
- Group Key
- ...

**Marketplace**
- Groups Count to Date
- Interest Rates Table
- Loans Closed Count
- ...

**Credit Profile**
- Amount Delinquent
- Bankcard Utilization
- Credit Grade
- ...

**Loan Performance**
- Cycle Number
- DPD (Date Past Due)
- Net Defaults
- ...
“the original 11”

- Amount Requested
- Bid Count *
- Borrower Rate
- Credit Grade
- Debt to Income Ratio
- Group Key
- Has an image
- Current delinquencies
- Delinquencies last 7 years
- Open credit lines
- Income

* Not used in Loans vs. unfunded listings classification
PCA shows separability
Variance and principle components
Hi and thank you for looking at my post. I currently own a small 3 employee business in Minnesota, I started the business about 3 years ago and it is really taking off. We currently have over $100,000 in inventory and are looking to hire more employees. I would like this loan to actually buy even more inventory and also just to have fun with Prosper and use it, I love lending people money on here, it is way more fun then the stock market. Our current sales are about $360,000/yr. but we are on track for $500,000 this next fiscal year.

A little info about myself: I am married to a wonderful woman and we have a baby on the way. We have a beautiful big brown Newfoundland named Tank and a Persian cat named Goo who hates me. My hobbies include playing hockey, flying small private planes, and building stuff around my house.

Thank you!
### Descriptions

<table>
<thead>
<tr>
<th>Loans</th>
<th>Listings</th>
<th>Difference</th>
<th>Words</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.44</td>
<td>0.58</td>
<td>0.14</td>
<td>cards and other</td>
</tr>
<tr>
<td>0.29</td>
<td>0.42</td>
<td>0.14</td>
<td>monthly expenses housing</td>
</tr>
<tr>
<td>0.44</td>
<td>0.58</td>
<td>0.14</td>
<td>clothing household expenses</td>
</tr>
<tr>
<td>0.41</td>
<td>0.54</td>
<td>0.13</td>
<td>and other loans</td>
</tr>
<tr>
<td>0.42</td>
<td>0.56</td>
<td>0.14</td>
<td>other expenses</td>
</tr>
<tr>
<td>0.29</td>
<td>0.43</td>
<td>0.14</td>
<td>expenses housing</td>
</tr>
<tr>
<td>0.44</td>
<td>0.57</td>
<td>0.13</td>
<td>clothing household</td>
</tr>
<tr>
<td>0.45</td>
<td>0.58</td>
<td>0.13</td>
<td>car expenses</td>
</tr>
</tbody>
</table>

### Titles

<table>
<thead>
<tr>
<th>Loans</th>
<th>Listings</th>
<th>Difference</th>
<th>Words</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.024</td>
<td>0.014</td>
<td>0.010</td>
<td>high interest credit</td>
</tr>
<tr>
<td>0.022</td>
<td>0.015</td>
<td>0.007</td>
<td>off high interest</td>
</tr>
<tr>
<td>0.021</td>
<td>0.016</td>
<td>0.005</td>
<td>credit card debt</td>
</tr>
<tr>
<td>0.009</td>
<td>0.004</td>
<td>0.005</td>
<td>interest credit card</td>
</tr>
<tr>
<td>0.062</td>
<td>0.04</td>
<td>0.022</td>
<td>credit card</td>
</tr>
<tr>
<td>0.049</td>
<td>0.032</td>
<td>0.017</td>
<td>high interest credit</td>
</tr>
<tr>
<td>0.015</td>
<td>0.003</td>
<td>0.012</td>
<td>in prosper</td>
</tr>
<tr>
<td>0.026</td>
<td>0.015</td>
<td>0.011</td>
<td>interest card</td>
</tr>
</tbody>
</table>
the 96
Floating Feature Search

- Linear Discriminant as evaluation function
- Lots of samples, lots of features:
  - 96 features
  - 300,000 listings
- ssssssssllllllloooooooowwwwwwwwww. Must:
  - limit the number of features – hence *forwards* floating search. Decreases optimality.
  - decrease the number of samples (increases bias)
Feature Crawl

Position in feature space vs. Several hours

- 87 features, 0.8
- 80 features, 0.802
- 34 features, 0.79875
- 23 features, 0.8
Awesomeness of Features

- Credit Grade
- "is a" in Member endorsements
- number of open credit lines
- "pay off" in Title

- Is Borrower Homeowner?
- Number of capital letters in title

Number of times feature was chosen vs. Which feature
Classification Performance

Linear Discriminant Classification

Not Awarded Loan
Awarded Loan

Credit Grade

Borrower Max. Rate

Amount Requested
Classification Overview

• Review of Methods
• Discussion of prior probability, implications when viewing results
• Summary of Results, Tables
• Classification Improvements
• Tying it Back to P2P Lending

• Detailed performance data is given on project website.
What We Did

Classification Procedure

FS11

FS96

4 Classifiers

Performance
Review of Methods

• LDA – pseudoinverse (mse)
• PCA+LDA – reduce dims, classify using some number of principal comp. (70%/30% 10F CV)
• SVM – map higher to higher dim space
  • Linear, 1 norm (smo) soft margin (slack) (70%/30%, no CV)
• Neural Networks – high degree of freedom (hidden nodes)
  • Feedforward, Tr:conjugate gradient descent. 1 HL {5,10,20 hidden nodes} 70%tr,15%val, 10%test

• Libraries: Matlab SVM, NN Tools
• Prior probability leads to classifiers that favor one class
• In comparing classifiers using stratified sampling FN rate is large contributor of error
• Results are given for both stratified sampling & equal priors
• PCA+LDA classification attempts to separate these distributions in a lower dim
Brief Summary of Results

• Listing\Loan (Stratified) best performing classifier
  • FS10, Neural Network (10 hidden nodes)
  • **14% error (prevalence of C1: 16.8%)**
    – 2000 samples, 20 Hidden, 75%Tr, 15%Val, 10%Test
  • Different prior probabilities of FS10 & FS96
    – Effect is that error is mostly FN approx. = prevalence of c1

• Listing\Loan (Equal) best performing classifier
  • FS10, FS96 Neural Network (20 hidden nodes)
  • **18%, 16% respectively**
    – 2000 samples, 20 Hidden, 75%Tr, 15%Val, 10%Test

• Default\No Default
  • FS11, FS96 Neural Network (20 hidden nodes)
  • **26%,15% respectively**
    – 2000 samples, 20 Hidden, 75%Tr, 15%Val, 10%Test
Lessons \ Future Improvements

- Neural Nets were a good match, surprising?
  - Not really, given a number of hidden nodes (degrees of freedom), arbitrarily complex decision boundaries can be found – great for high dimension feature vectors.

- Effect of adding additional features
  - For any method of classification, data suggests additional dimensions improve accuracy, complexity not worth the effort. We are talking about ~4% less error.

- Satisficing – LDA+PCA good enough
  - No matter what method was used results ~84-79% correct

- Real world data != equal priors
  - Feature search should seek to minimize FP, FN – better separation, more realistic classifiers for Loan/No Loan
Take Home Message

• What is all this really good for anyways?
  • Designed several classifiers performing > 80% accuracy (that’s great but…)
  • Goal is not to make the world’s best performing classifier, rather – the data can be classified. (clustering-> classification -> intuitive models)

• A Strategy for Borrowers?
  • We demonstrated that there are features that can separate the data, what is your strategy to improve chances…
    • Classification results not quite satisfying and tractable
    • Coco & Ernesto build on these results – demonstrate models that make intuitive sense

• A Strategy for Lenders – an Investment Tool?
  • We demonstrated to a reasonable accuracy that features indicative of defaulting exist. Is this better than intuition? (a machine classifier doesn’t say you should get a loan because you got divorced, experienced a disaster, etc)
  • Problems with this strategy – Someone of dubious repayment potential gets a loan. She then repays because she won the lottery. Outside events not taken into account.
  • A temporal analysis to examine loan performance is required. “Now that you have a loan lets see what you do.” Stay tuned for HMMs
### Loan / No Loan Summary

<table>
<thead>
<tr>
<th>Method</th>
<th>FS96 E</th>
<th>FS10 E</th>
<th>#FS9 6</th>
<th>#FS1 0</th>
<th>Ratio FS96</th>
<th>Ratio FS10</th>
<th>Trl(Val)/Te</th>
<th>CV FS96</th>
<th>CV FS11</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM (All)</td>
<td>0.085</td>
<td>0.157</td>
<td>2000</td>
<td>2000</td>
<td>0.085</td>
<td>0.168</td>
<td>70/30</td>
<td>None</td>
<td>None</td>
</tr>
<tr>
<td>SVM TOP 10</td>
<td>0.098</td>
<td>null</td>
<td>2000</td>
<td>2000</td>
<td>0.085</td>
<td>0.168</td>
<td>70/30</td>
<td>None</td>
<td>None</td>
</tr>
<tr>
<td>PCA+LDA (10 P.C.)</td>
<td>0.259</td>
<td>0.223</td>
<td>5000</td>
<td>5000</td>
<td>0.085</td>
<td>0.168</td>
<td>70/30</td>
<td>10 Fold CV</td>
<td>10 Fold CV</td>
</tr>
<tr>
<td>NN ALL (20 Hidden)</td>
<td>0.085</td>
<td>0.143</td>
<td>2000</td>
<td>2000</td>
<td>0.085</td>
<td>0.168</td>
<td>75/15/10</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>NN ALL (10 Hidden)</td>
<td>0.085</td>
<td>0.141</td>
<td>2000</td>
<td>2000</td>
<td>0.085</td>
<td>0.168</td>
<td>75/15/10</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>NN ALL (5 Hidden)</td>
<td>0.085</td>
<td>0.145</td>
<td>2000</td>
<td>2000</td>
<td>0.085</td>
<td>0.168</td>
<td>75/15/10</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>LDA 30</td>
<td>0.085</td>
<td>null</td>
<td>10000</td>
<td>null</td>
<td>0.085</td>
<td>null</td>
<td>70/30</td>
<td>None</td>
<td>null</td>
</tr>
<tr>
<td>LDA 10</td>
<td>0.085</td>
<td>0.141</td>
<td>10000</td>
<td>10000</td>
<td>0.085</td>
<td>0.168</td>
<td>70/30</td>
<td>None</td>
<td>10 Fold CV</td>
</tr>
<tr>
<td>LDA 5</td>
<td>0.085</td>
<td>null</td>
<td>10000</td>
<td>null</td>
<td>0.085</td>
<td>null</td>
<td>70/30</td>
<td>None</td>
<td>null</td>
</tr>
</tbody>
</table>

Note: while the error in this case is high, the FN classification is better due to pca dim reduction.

### Stratified sampling

<table>
<thead>
<tr>
<th>Method</th>
<th>FS96 E</th>
<th>FS10 E</th>
<th>#FS9 6</th>
<th>#FS1 0</th>
<th>Ratio FS96</th>
<th>Ratio FS10</th>
<th>Trl(Val)/Te</th>
<th>CV FS96</th>
<th>CV FS11</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM (All)</td>
<td>0.192</td>
<td>0.230</td>
<td>2000</td>
<td>2000</td>
<td>0.500</td>
<td>0.500</td>
<td>70/30</td>
<td>None</td>
<td>None</td>
</tr>
<tr>
<td>SVM TOP 10</td>
<td>0.212</td>
<td>null</td>
<td>2000</td>
<td>2000</td>
<td>0.500</td>
<td>0.500</td>
<td>70/30</td>
<td>None</td>
<td>None</td>
</tr>
<tr>
<td>PCA+LDA (10 P.C.)</td>
<td>0.259</td>
<td>0.223</td>
<td>5000</td>
<td>5000</td>
<td>0.500</td>
<td>0.500</td>
<td>70/30</td>
<td>10 Fold CV</td>
<td>10 Fold CV</td>
</tr>
<tr>
<td>NN ALL (20 Hidden)</td>
<td>0.160</td>
<td>0.185</td>
<td>2000</td>
<td>2000</td>
<td>0.500</td>
<td>0.500</td>
<td>75/15/10</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>NN ALL (10 Hidden)</td>
<td>0.183</td>
<td>0.191</td>
<td>2000</td>
<td>2000</td>
<td>0.500</td>
<td>0.500</td>
<td>75/15/10</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>NN ALL (5 Hidden)</td>
<td>0.212</td>
<td>0.200</td>
<td>2000</td>
<td>2000</td>
<td>0.500</td>
<td>0.500</td>
<td>75/15/10</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>LDA 30</td>
<td>0.213</td>
<td>null</td>
<td>10000</td>
<td>null</td>
<td>0.500</td>
<td>null</td>
<td>70/30</td>
<td>None</td>
<td>null</td>
</tr>
<tr>
<td>LDA 10</td>
<td>0.259</td>
<td>0.220</td>
<td>10000</td>
<td>10000</td>
<td>0.500</td>
<td>0.500</td>
<td>70/30</td>
<td>None</td>
<td>10 Fold CV</td>
</tr>
<tr>
<td>LDA 5</td>
<td>0.224</td>
<td>null</td>
<td>10000</td>
<td>null</td>
<td>0.500</td>
<td>null</td>
<td>70/30</td>
<td>None</td>
<td>null</td>
</tr>
</tbody>
</table>

### Equal Sampling

<table>
<thead>
<tr>
<th>Method</th>
<th>FS96 E</th>
<th>FS10 E</th>
<th>#FS9 6</th>
<th>#FS1 0</th>
<th>Ratio FS96</th>
<th>Ratio FS10</th>
<th>Trl(Val)/Te</th>
<th>CV FS96</th>
<th>CV FS11</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM (All)</td>
<td>0.192</td>
<td>0.230</td>
<td>2000</td>
<td>2000</td>
<td>0.500</td>
<td>0.500</td>
<td>70/30</td>
<td>None</td>
<td>None</td>
</tr>
<tr>
<td>SVM TOP 10</td>
<td>0.212</td>
<td>null</td>
<td>2000</td>
<td>2000</td>
<td>0.500</td>
<td>0.500</td>
<td>70/30</td>
<td>None</td>
<td>None</td>
</tr>
<tr>
<td>PCA+LDA (10 P.C.)</td>
<td>0.259</td>
<td>0.223</td>
<td>5000</td>
<td>5000</td>
<td>0.500</td>
<td>0.500</td>
<td>70/30</td>
<td>10 Fold CV</td>
<td>10 Fold CV</td>
</tr>
<tr>
<td>NN ALL (20 Hidden)</td>
<td>0.160</td>
<td>0.185</td>
<td>2000</td>
<td>2000</td>
<td>0.500</td>
<td>0.500</td>
<td>75/15/10</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>NN ALL (10 Hidden)</td>
<td>0.183</td>
<td>0.191</td>
<td>2000</td>
<td>2000</td>
<td>0.500</td>
<td>0.500</td>
<td>75/15/10</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>NN ALL (5 Hidden)</td>
<td>0.212</td>
<td>0.200</td>
<td>2000</td>
<td>2000</td>
<td>0.500</td>
<td>0.500</td>
<td>75/15/10</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>LDA 30</td>
<td>0.213</td>
<td>null</td>
<td>10000</td>
<td>null</td>
<td>0.500</td>
<td>null</td>
<td>70/30</td>
<td>None</td>
<td>null</td>
</tr>
<tr>
<td>LDA 10</td>
<td>0.259</td>
<td>0.220</td>
<td>10000</td>
<td>10000</td>
<td>0.500</td>
<td>0.500</td>
<td>70/30</td>
<td>None</td>
<td>10 Fold CV</td>
</tr>
<tr>
<td>LDA 5</td>
<td>0.224</td>
<td>null</td>
<td>10000</td>
<td>null</td>
<td>0.500</td>
<td>null</td>
<td>70/30</td>
<td>None</td>
<td>null</td>
</tr>
</tbody>
</table>
## Default / No Default Summary

<table>
<thead>
<tr>
<th>Method</th>
<th>FS96 E</th>
<th>FS11 E</th>
<th>#FS9 6</th>
<th>#FS1 1</th>
<th>Ratio FS96</th>
<th>Ratio FS11</th>
<th>Tr/ (Val)/Te</th>
<th>CV FS96</th>
<th>CV FS11</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM (All)</td>
<td>0.190</td>
<td>0.270</td>
<td>2000</td>
<td>2000</td>
<td>0.500</td>
<td>0.500</td>
<td>70/30</td>
<td>None</td>
<td>None</td>
</tr>
<tr>
<td>SVM TOP 10</td>
<td>0.210</td>
<td>null</td>
<td>2000</td>
<td>2000</td>
<td>0.500</td>
<td>0.500</td>
<td>70/30</td>
<td>None</td>
<td>None</td>
</tr>
<tr>
<td>PCA+LDA (10 P.C.)</td>
<td>0.197</td>
<td>0.250</td>
<td>2000</td>
<td>2000</td>
<td>0.500</td>
<td>0.500</td>
<td>70/30</td>
<td>10 Fold CV</td>
<td>10 Fold CV</td>
</tr>
<tr>
<td>NN ALL (20 Hidden)</td>
<td>0.152</td>
<td>0.264</td>
<td>2000</td>
<td>2000</td>
<td>0.500</td>
<td>0.500</td>
<td>75/15/10</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>NN ALL (10 Hidden)</td>
<td>0.156</td>
<td>0.271</td>
<td>2000</td>
<td>2000</td>
<td>0.500</td>
<td>0.500</td>
<td>75/15/10</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>NN ALL (5 Hidden)</td>
<td>0.164</td>
<td>0.273</td>
<td>2000</td>
<td>2000</td>
<td>0.500</td>
<td>0.500</td>
<td>75/15/10</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>LDA 30</td>
<td>0.273</td>
<td>null</td>
<td>1000</td>
<td>null</td>
<td>0.500</td>
<td>null</td>
<td>70/30</td>
<td>None</td>
<td>null</td>
</tr>
<tr>
<td>LDA 11</td>
<td>0.257</td>
<td>0.240</td>
<td>1000</td>
<td>2000</td>
<td>0.050</td>
<td>0.500</td>
<td>70/30</td>
<td>None</td>
<td>10 Fold CV</td>
</tr>
<tr>
<td>LDA 5</td>
<td>0.248</td>
<td>null</td>
<td>1000</td>
<td>null</td>
<td>null</td>
<td>null</td>
<td>70/30</td>
<td>None</td>
<td>null</td>
</tr>
</tbody>
</table>
Bayesian Network

• Nine Quantized Features from Floating Selection Set:

A. Amount Delinquent (Low, High)
B. Open Credit Lines (Low, Med, High)
C. Amount Requested (Low, Moderate, High, Very High)
D. Borrower’s Max Rate (Low, Moderate, High, Very High)
E. Credit Grade (Poor, Average, Good, Very good)
F. Debt to Income Ratio (Low, Med, High)
G. ‘Good Candidate’ (True, False)
H. Funding Option (True, False)
I. Endorsement (True, False)
Structure Learning

• Methods:
  – Exhaustive Search: PC Algorithm
  – Score-Based: MCMC; K2, Greedy Search

• Challenges:
  – 10 Nodes = $4.2 \times 10^{18}$ Directed Acyclic Graphs!!!!!!!
  – PC algorithm… Overflow!!!
  – Overfitting??
Complete Graph
Building Other Models

- Models:
  - Naïve Bayes Classifier
  - Belief Structures
  - Noisy Functional Dependence Models

- Parameter Estimation (complete data set)
  - *Batch Learning*: MLE & Bayesian Estimation
    (Maximum a posteriori parameters)
  - *MAP decision rule* (classification)
Naïve Bayes Classification

Loan

A
B
C
D
E
F
G
H
I
Noisy Functional Dependence

- Open Credits
- Debt to Income
- Amount Requested
- Max Rate
- Amount Delinquent
- Credit Grade
- Funding Opt.
- ‘Good Candidate’
- Loan Risk
- Member support
- Endorsement

Profile

Loan
<table>
<thead>
<tr>
<th>Bayes Net Model</th>
<th>BIC score (x 10^5)</th>
<th>Clsf. Performance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Learned Structure (MCMC)</td>
<td>-1.33</td>
<td>0.76</td>
</tr>
<tr>
<td>Learned Structure (Chow-Liu &amp; K2)</td>
<td>-1.34</td>
<td>0.7525</td>
</tr>
<tr>
<td>Naïve Bayes Net</td>
<td>-1.83</td>
<td>0.7580</td>
</tr>
<tr>
<td>Belief structure</td>
<td>-1.42</td>
<td>0.5620</td>
</tr>
</tbody>
</table>
Decision Trees
Decision Trees: questions

- BORROWER: will my loan get funded?
  - (how much should I borrow? what interest rate should I set?)

- LENDER: if I fund this loan, will I be paid back?
  - (what features best predict default? which loans should I fund?)
Decision Trees: methodology

• FEATURES: experimented with various sets
  
  • Greedy 11 (eliminated #bids)
  
  • sets of 2 4 6 8 10 features
  
• NODE SPLIT THRESHOLD: 2 - 6400
  
  • minimize Gini impurity
    
    \[ I_G(i) = 1 - \sum_{j=1}^{m} f(i, j)^2 = \sum_{j \neq k} f(i, j)f(i, k) \]
    
  probability \( i \) belongs to class \( j \) to zero when all samples part of single target category

• PRIORS: tried with / without prior probabilities [13% loan, 87% no loan]
Decision Tree: analysis

Variables:
- FEATURE SET
- NODE SPLIT threshold

Tests:
- SENSITIVITY: reserved 10%, 10 iterations
- ERROR RATES: total, FP, FN

Pick best feature sets & split threshold to:
minimize variance across iterations
minimize error
now entering: TINY FONT TERRITORY
is my loan likely to get funded?

- B credit
- no delinq
- DTI 10%
- 11% interest
- $1500 loan
is my loan likely to get funded?

- B credit
- no delinq
- DTI 10%
- 11% interest
- $1500 loan
is my loan likely to get funded?

- B credit
- no delinq
- DTI 10%
- 11% interest
- $1500 loan
is my loan likely to get funded?

- **B credit**
- no delinquency
- DTI 10%
- 11% interest
- $1500 loan

2000 split, 13/87 pp
is my loan likely to get funded?

- B credit
- no delinq
- DTI 10%
- 11% interest
- $1500 loan
is my loan likely to get funded?

- B credit
- no delinq
- DTI 10%
- 11%
  interest
- $1500 loan
is my loan likely to get funded?

- B credit
- no delinq
- DTI 10%
- 11% interest
- $1500 loan
is my loan likely to get funded?

- B credit
- no delinq
- DTI 10%
- 11% interest
- $1500 loan
is my loan likely to get funded?

- B credit
- no delinq
- DTI 10%
- 11% interest
- $1500 loan

cre <580

del < 0.5
dti <.405
int <.249
amt < 9999.5
cre <660
int <.076

2000 split, 13/87 pp
is my loan likely to get funded?

- B credit
- no delinq
- DTI 10%
- 11% interest
- $1500 loan

2000 split, 13/87 pp
is my loan likely to get funded?

- B credit
- no delinq
- DTI 10%
- 11% interest
- $1500 loan

but not with C credit!
is my loan likely to get funded?

- B credit
- no delinq
- DTI 10%
- 11% interest
- $1500 loan

but not with C credit!
priors matter!
(same profile, without priors, predicts no loan)
DT could be used to help borrowers set loan amount, increase loan conversion for prosper

C credit, DTI = 10%, 1 current delinquency, needs $6000

tree predicts (borrowers with same profile)

56% no loan
average amt listed: $4625

44% loan
average amt listed: $3700

advice: request lower amount
(analysis of reposted loans)

1500 members reposted at same amt

4037 sets of reposts (same person, same amt)

223 eventually completed

only 5 (0.3%) change loan amount
analogous process for default / paid

- Priors closer to 50/50

- Lender can use DT to identify conditional probability of default given Profile X

- Important for Prosper: keep tabs on loans with high default risk
optimal pruning level
(default paid tree, 6 features)

overfit

not enough granularity
as features are added, error rate down
(loan / no loan, 200 split threshold)
adding features, reducing split level decreases error, sensitivity

![Graph showing sensitivity and error by number of features](image)

- 10 iterations
- 10% holdout
FP down, FN up as node split threshold increases (loan no loan, 6 features)
Loan Performance HMMs

current  late  defaulted
current  late  defaulted
current  late  defaulted
HMM Performance

- Training 70%, testing 30% of loan performance data
  - Paid: 4365 sequences
  - Default: 3793 sequences
  - All sequences of varying length
- Model was verified and then tested
HMM Results

• Why build a model for someone who will default?
  – Short term \ long term visibility of loan performance is important
  – Default behavior potentially mimics fraud
  – A lender may be well aware of loan performance but what about Prosper?
    • Improved customer service – easier to monitor high risk loans, early contact of collection agency
    • Default Performance may mimic fraud (Prosper has problems with this).
Improving HMM Performance

- Retrain specifying pi, currently start at state 1
- Create a single model of financial health
  - Train model using both paid / default data
  - Use Viterbi algorithm to estimate “proximity” to hidden state that best characterizes defaults (easy)
- Hierarchical HMM (complex)
  - Advantage is that HMMs can emit sequences of observations
    - A way to reduce error in early stages?
  - Reading
not all groups are created equal: refining social features

- Some groups have much higher funding rates than others (queried tier-2 description, sort by category, % members funded by individual group)

<table>
<thead>
<tr>
<th>highest % funded</th>
<th>best funded, popular groups</th>
<th>lowest % funded</th>
</tr>
</thead>
<tbody>
<tr>
<td>Albuquerque</td>
<td>Florida</td>
<td>Adoption Agencies</td>
</tr>
<tr>
<td>Aviation</td>
<td>Extended Families</td>
<td>Air Quality</td>
</tr>
<tr>
<td>Greece</td>
<td>Research &amp; Analysis</td>
<td>Amateur</td>
</tr>
<tr>
<td>Oil &amp; Gas</td>
<td>Massachusetts</td>
<td>Beading</td>
</tr>
<tr>
<td>Opthamology</td>
<td>Travel</td>
<td>Big East Conference</td>
</tr>
<tr>
<td>Poverty Relief</td>
<td>Accounting</td>
<td>Chemical</td>
</tr>
<tr>
<td>Rhode Island</td>
<td>Pennsylvania</td>
<td>Construction</td>
</tr>
<tr>
<td>Rugby</td>
<td>Software</td>
<td>Deist</td>
</tr>
<tr>
<td>Seattle</td>
<td>Financial Planning</td>
<td>DJs</td>
</tr>
<tr>
<td>Space</td>
<td>Mortgage</td>
<td>Equipment &amp; Tools</td>
</tr>
<tr>
<td>Theatre</td>
<td>Small &amp; Medium Business</td>
<td>Estimating</td>
</tr>
<tr>
<td>West Virginia</td>
<td>Investment Management</td>
<td>Fiction</td>
</tr>
<tr>
<td>Veterinary</td>
<td>Family Owned</td>
<td>Gambling</td>
</tr>
<tr>
<td></td>
<td>Virginia</td>
<td>Glass</td>
</tr>
<tr>
<td></td>
<td>Financial Consultants</td>
<td>Gliding</td>
</tr>
<tr>
<td></td>
<td>Education &amp; Training</td>
<td>Jewelry</td>
</tr>
<tr>
<td></td>
<td>Large Families</td>
<td>Kentucky</td>
</tr>
<tr>
<td></td>
<td>Catholic</td>
<td>Kung Fu</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Law Firms</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>lowest % funded</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Adoption Agencies</td>
<td>Air Quality</td>
<td>Oceania</td>
</tr>
<tr>
<td>Amateur</td>
<td>Beading</td>
<td>Poetry</td>
</tr>
<tr>
<td>Big East Conference</td>
<td>Chemical</td>
<td>Printmaking</td>
</tr>
<tr>
<td>Chemical</td>
<td>Construction</td>
<td>Recycling</td>
</tr>
<tr>
<td>Construction</td>
<td>Deist</td>
<td>Refugees</td>
</tr>
<tr>
<td>Deist</td>
<td>DJs</td>
<td>Republicans</td>
</tr>
<tr>
<td>DJs</td>
<td>Equipment &amp; Tools</td>
<td>Security</td>
</tr>
<tr>
<td>Equipment &amp; Tools</td>
<td>Estimating</td>
<td>Senior Citizens</td>
</tr>
<tr>
<td>Estimating</td>
<td>Fiction</td>
<td>Sporting Goods</td>
</tr>
<tr>
<td>Fiction</td>
<td>Gambling</td>
<td>Sporting Goods</td>
</tr>
<tr>
<td>Gambling</td>
<td>Glass</td>
<td>Structural</td>
</tr>
<tr>
<td>Glass</td>
<td>Gliding</td>
<td>Surety Bonds</td>
</tr>
<tr>
<td>Gliding</td>
<td>Jewelry</td>
<td>Symphony</td>
</tr>
<tr>
<td>Jewelry</td>
<td>Kentucky</td>
<td>Thailand</td>
</tr>
<tr>
<td>Kentucky</td>
<td>Kung Fu</td>
<td>Trading Cards</td>
</tr>
<tr>
<td>Kung Fu</td>
<td>Law Firms</td>
<td>United Kingdom</td>
</tr>
<tr>
<td>Law Firms</td>
<td></td>
<td>Utilities</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Yoga</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Youth</td>
</tr>
</tbody>
</table>
Using Amazon Mechanical Turk to classify images
a human classification experiment

- Unlike banks, prosper lenders can weigh more than “just the numbers”

- Banks seek ROI; prosper lenders may have other motives (e.g. social good)

- Prosper lenders lack complex risk algorithms of banks

- Many borrowers may meet a lender’s baseline criteria (e.g. FICO > 600) ... social criteria and profile assessment needed to decide how to allocate funds

- Holistic assessment of borrower profile: necessary and natural
does a borrower seem “trustworthy”?

• Goal 1: image classification

• Goal 2: assessment of “trustworthiness”

• Does “trustworthiness” correlate with getting a loan?

• Here, only pilot of methodology and analysis

• Follow-up could use humans to train classifier or create feature vector
amazon mechanical turk

Tag this image

Guidelines:

- Check the best description of what's in the image
- Check the best answer to the question: "Does this person (or the person who posted the image) look trustworthy?"
- Your answer to the first question will be calibrated against others to ensure correct tagging

Image:

What's in this picture?

- Adult
- Child/children
- Friends/Family
- Landscape
- Animal
- House
- Vehicle
- Other

Does this person (or the person who posted the image) look trustworthy?  □ Trustworthy  □ Untrustworthy
data collection

• 200 images x 2 questions x 3 workers / image (used to check for consistency)

• 50% images from unfulfilled listings; 50% from paid-off loans
consistency was good, especially for categorization

- CATEGORIES: confusion from label choice; 9.5% between children/family

  89.5%  9.5%  1%

- (1% disagreed on how to categorize e.g. a vehicle + people)

- Multiple opinions good as fuzzy categorization?
trust rating requires clarification

- TRUST: 11% disagreement, both contextual and subjective

- Lack of context

- Blurry photo, real distrust?
categories and trust: 71% of photos have people

categories

trust
no correlation between getting a loan and trust tag

- an image tagged “untrustworthy” was just as likely to have received & paid a loan as to have listed with no loan (no statistical difference)

- would adding contextualization (listing description) or refining the question phrasing help classification?

- Research question: how independent is judgment of “trustworthiness” from the stories built from contextual information (credit score, loan purpose), especially for quick (~8 seconds / photo) decisions?
human classification: analysis

- human-augmented classification can work: consistency was high

- experiment design is important: vague questions yield vague results

- future work could collect larger sample; use as a feature vector

- also, text / spelling analysis