Pseudo-Social Network Targeting from Consumer Transaction Data

Martens, David and Provost, Foster

September 2011
Abstract

“This design science paper presents a method for targeting consumers based on a 'pseudo-social network' (PSN): consumers are linked if they transfer money to the same entities. A marketer can target those individuals that are strongly connected to key individuals...”
Introduction

● The goal of consumer targeting* is to target the “right” audience rather than just the mass audience.
● Efficient and accurate techniques of targeting customers can ultimately increase sales.

“Targeting in marketing is a strategy that breaks a large market into smaller segments to concentrate on a specific group of customers within that audience.”
Background

Predictive modelling
Uses statistics to predict outcomes. Is an email ham or spam?

Socio-demographic variables
- Demographics, psychographics, company activity, etc.
- State-of-the-art according to this paper

Social network targeting
- Create an explicitly represented social network among consumers
- Select or rank consumers to target based (in part) on their proximity in the network to selected individuals of interest, such as existing customers
- 3 to 5 times more effective than predictive modeling for targeting telecom offers (Hill et al. 2006, Aral et al. 2009)
New approach

Pseudo-social network (PSN)

- A graph representation of consumers and institutions
- Two consumers are linked if they transfer money to the same people or institutions
  - Includes transfers to the consumer such as paychecks and transfers to institutions such as paying for a service
- *Pseudo* because the network does not represent a true social network
- Strong connections in the PSN -> Useful similarity between consumers
- Privacy-friendly, since it is not necessary to know the identities of the connected consumers
Design of the PSN

- Built on financial transaction data
- Similarity - “two consumers are similar if they make payments to the same entity, or receive payments from the same entity, and are more similar the more such connections they share and the stronger the connections.”
From money transfers to a PSN

- Payment receivers such as Amazon and EnergyInc provide little information on the similarity between consumers
From money transfer data to a PSN scoring

Payment transactions are converted to a payment receiver matrix

- Known buyers vs. non-known buyers

<table>
<thead>
<tr>
<th>pr</th>
<th>Consumers</th>
<th>NC(pr)</th>
<th>NB(pr)</th>
<th>ICF(pr)</th>
</tr>
</thead>
<tbody>
<tr>
<td>LittleBookStore</td>
<td>a b c</td>
<td>3</td>
<td>2</td>
<td>1.52</td>
</tr>
<tr>
<td>DeliC</td>
<td>a d e</td>
<td>3</td>
<td>0</td>
<td>1.52</td>
</tr>
<tr>
<td>Amazon</td>
<td>f g h i</td>
<td>4</td>
<td>1</td>
<td>1.40</td>
</tr>
<tr>
<td>EnergyInc</td>
<td>b c d e f g</td>
<td>6</td>
<td>3</td>
<td>1.22</td>
</tr>
</tbody>
</table>

Table 1: Example: from transaction payment ($pr$) to PSN. The known buyers among the consumers are denoted in boldface.
From money transfer data to a PSN scoring

Similarity scores

- Ranking of consumers for targeting
- Inverse Consumer Frequency (ICF) \( \sim \) *Inverse Document Frequency*

\[ nc = \text{number of consumers} \]
\[ npr = \text{number of payment receivers} \]
\[ NC(pr) = \text{number of unique consumers having made a payment to } pr \]
\[ NB(pr) = \text{number of unique known buyers having made a payment to } pr \]
\[ B(x,pr) = 1 \text{ if consumer } x \text{ made a payment to } pr \]
\[ = 0 \text{ if consumer } x \text{ did not make a payment to } pr \]

\[ ICF(PR) = \log_{10} \left( \frac{nc}{NC(pr)} \right) \quad (1) \]

\[ \text{Score}(x) = \sum_{i=1}^{npr} \left( \frac{NB(pr)}{NC(pr)} \times ICF(pr_i) \right) \times B(x,pr_i) \quad (2) \]
Example of calculation

<table>
<thead>
<tr>
<th>pr</th>
<th>Consumers</th>
<th>NC(pr)</th>
<th>NB(pr)</th>
<th>ICF(pr)</th>
</tr>
</thead>
<tbody>
<tr>
<td>LittleBookStore</td>
<td>a b c</td>
<td>3</td>
<td>2</td>
<td>1.52</td>
</tr>
<tr>
<td>DeliC</td>
<td>a d e</td>
<td>3</td>
<td>0</td>
<td>1.52</td>
</tr>
<tr>
<td>Amazon</td>
<td>f g h i</td>
<td>4</td>
<td>1</td>
<td>1.40</td>
</tr>
<tr>
<td>EnergyInc</td>
<td>b c d e f g</td>
<td>6</td>
<td>3</td>
<td>1.22</td>
</tr>
</tbody>
</table>

\[ \text{Score}(a) = \text{Score}_{\text{LittleBookStore}} + \text{Score}_{\text{DeliC}} \]
\[ = \left( \frac{2}{3} \times 1.52 \right) + \left( \frac{0}{3} \times 1.52 \right) \]
\[ = 1.01 \quad (3) \]
\[ \text{Score}(d) = \text{Score}_{\text{DeliC}} + \text{Score}_{\text{EnergyInc}} \]
\[ = \left( \frac{0}{3} \times 1.52 \right) + \left( \frac{3}{6} \times 1.22 \right) \]
\[ = 0.61 \quad (4) \]

Score(e) = 0.61  
Score(f) = 0.96  
Score(h) = 0.35  
Score(i) = 0.35

In this example the highest score is obtained by consumer a, so we should offer a the product that the known buyers purchased.
The data

- 5 million (debit) transactions
- 1.2 million customers
- 3.2 million payment receivers (PRs)
- Data from a 11 month period
Experimental setup

The PSN model was compared against random targeting and four competing models

1. Linear support vector machine (SVM) $\rightarrow SD_{\text{lin}}$
2. Non-linear SVM using the radial basis function (RBF) kernel $\rightarrow SD_{\text{RBF}}$
3. PSN + linear SVM $\rightarrow$ PSN + $SD_{\text{lin}}$
4. PSN + non-linear SVM $\rightarrow$ PSN + $SD_{\text{RBF}}$
Experimental Results

- **AUC (Area Under the ROC Curve)**
  - AUC is the probability that the model ranks a random positive example more highly than a random negative example.
  - Typically used for evaluating performance of binary classifiers
    - scale-invariant: rates how well predictions are ranked, instead of their abs values
    - does not evaluate based on a specific threshold

- **Lift**
  - The ratio of the target response over the average response
  - “For example, if 5% of the consumers are known buyers, and we are able to identify a segment (e.g., the 1% of consumers with the highest score) where the predicted response is 15%, a lift of 3 is obtained”

- **Lift > AUC**
  - Only neighbours of known buyers are provided with a score -> Most of the network remains unscored
## Experimental Results

<table>
<thead>
<tr>
<th></th>
<th>AUC</th>
<th>Lift 1%</th>
<th>Lift 5%</th>
<th>Lift 10%</th>
</tr>
</thead>
<tbody>
<tr>
<td>PSN</td>
<td>63.9</td>
<td>14.9</td>
<td>4.1</td>
<td>2.6</td>
</tr>
<tr>
<td>$SD_{lin}$</td>
<td>75.5</td>
<td>4.9</td>
<td>3.9</td>
<td>3.3</td>
</tr>
<tr>
<td>PSN + $SD_{lin}$</td>
<td><strong>78.2</strong></td>
<td><strong>12.7</strong></td>
<td><strong>5.4</strong></td>
<td><strong>4.0</strong></td>
</tr>
<tr>
<td>$SD_{RBF}$</td>
<td>77.6</td>
<td>4.8</td>
<td>4.3</td>
<td>3.6</td>
</tr>
<tr>
<td>PSN + $SD_{RBF}$</td>
<td>78.2</td>
<td>5.8</td>
<td>4.6</td>
<td>3.8</td>
</tr>
</tbody>
</table>

**Table 2** Results product 1 - 80% training data.

<table>
<thead>
<tr>
<th></th>
<th>AUC</th>
<th>Lift 1%</th>
<th>Lift 5%</th>
<th>Lift 10%</th>
</tr>
</thead>
<tbody>
<tr>
<td>PSN</td>
<td>71.7</td>
<td>31.8</td>
<td>7.6</td>
<td>4.4</td>
</tr>
<tr>
<td>$SD_{lin}$</td>
<td>86.6</td>
<td>10.1</td>
<td>7.8</td>
<td>6.0</td>
</tr>
<tr>
<td>PSN + $SD_{lin}$</td>
<td><strong>89.0</strong></td>
<td><strong>18.2</strong></td>
<td><strong>9.7</strong></td>
<td><strong>6.7</strong></td>
</tr>
<tr>
<td>$SD_{RBF}$</td>
<td>87.3</td>
<td>10.2</td>
<td>7.8</td>
<td>6.2</td>
</tr>
<tr>
<td>PSN + $SD_{RBF}$</td>
<td>87.4</td>
<td>10.6</td>
<td>7.8</td>
<td>6.2</td>
</tr>
</tbody>
</table>

**Table 3** Results product 2 - 80% training data.
Conclusions

- The major contribution is the demonstration of a new method for targeting customers using a PSN built from financial transaction data
- The combined model of PSN and SD is overall the most efficient, as the PSN and SD capture complimentary information
- More data would yield better results (indicates potential higher lift) for the PSN models as the AUC increases with higher % of training data, unlike for the SD$_{lin}$-model
- Another impact of their results is that it shows the concept of PSN based on transaction data can easily be applied to other business problems, such as estimating creditworthiness or discovering money laundering