Modeling Online Browsing and Path Analysis Using Clickstream Data

Alan Montgomery
Associate Professor
Carnegie Mellon University
Graduate School of Industrial Administration

With Shibo Li, Kannan Srinivasan, John Liechty

e-mail: alan.montgomery@cmu.edu
web: http://www.andrew.cmu.edu/user/alm3

JSM Meetings, 4 August 2003

Problem Overview

- Path Analysis
  - Infer consumer's purchase orientation from the navigation path
  - Impact of web design (i.e. hypertext links)
  - Impact of ads, price information, and free shipping promotion
- Purchase Conversion
  - Predict conversion dynamically
  - When and how to target online consumers

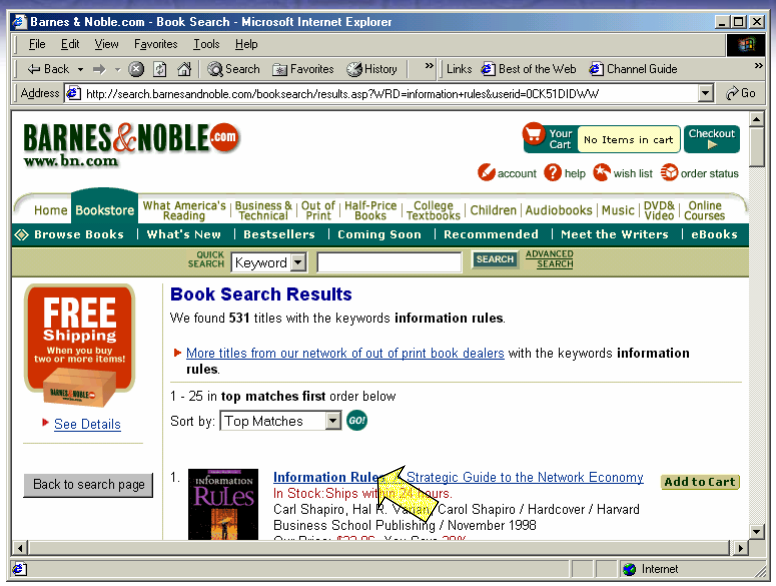
Outline

- Clickstream
- Data
- Model
- Results
- Purchase Conversion
- Conclusions

Clickstream

What is clickstream data?

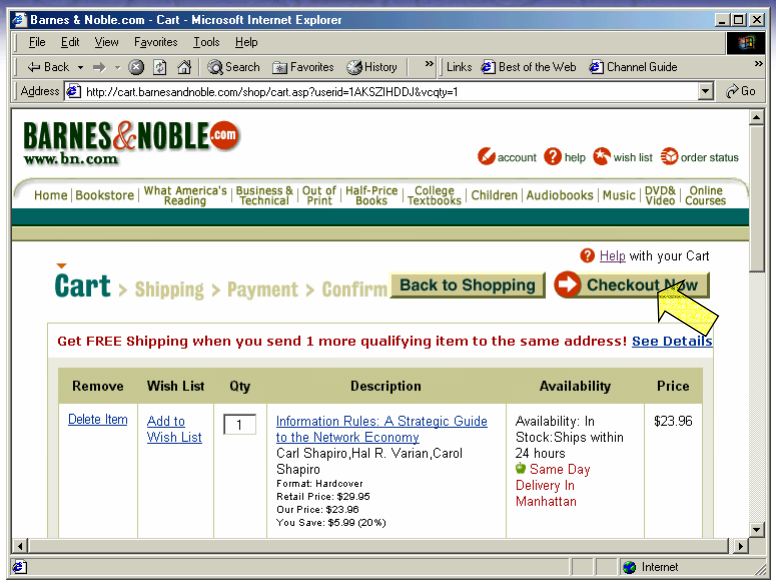
- A record of an individual's movement through time at a web site
- Contains information about:
  - Time
  - URL content
  - User's machine
  - Previous URL viewed
  - Browser type


Barnes & Noble.com - Book Search - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites History Links Best of the Web Channel Guide
Address http://search.barnesandnoble.com/booksearch/results.asp?WRD=information+rules&userid=0CK51DIDWW
Go
BARNES&NOBLE.com
www.bn.com
Your Cart No Items in cart Checkout
account help wish list order status
Home Bookstore What America's Reading Business & Technical Out of Print Half-Price Books College Textbooks Children Audiobooks Music DVD& Video Online Courses
Browse Books What's New Bestsellers Coming Soon Recommended Meet the Writers eBooks
QUICK SEARCH Keyword SEARCH ADVANCED SEARCH
FREE Shipping When you buy two or more items!
See Details
Back to search page
Book Search Results
We found 531 titles with the keywords information rules.
More titles from our network of out of print book dealers with the keywords information rules.
1 - 25 in top matches first order below
Sort by: Top Matches GO!
1.
Information Rules: A Strategic Guide to the Network Economy
In Stock:Ships within 24 hours.
Carl Shapiro, Hal R. Varian / Carol Shapiro / Hardcover / Harvard Business School Publishing / November 1998
Add to Cart
Internet

7


Barnes & Noble.com - Information Rules: A Strategic Guide to the Network Economy - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites History Links Best of the Web Channel Guide
Address http://search.barnesandnoble.com/booksearch/isbninquiry.asp?userid=1AKSZIHDDJ&isbn=087584863X
Go
BARNES&NOBLE.com
www.bn.com
Your Cart No Items in cart Checkout
account help wish list order status
Home Bookstore What America's Reading Business & Technical Out of Print Half-Price Books College Textbooks Children Audiobooks Music DVD& Video Online Courses
Browse Books What's New Bestsellers Coming Soon Recommended Meet the Writers eBooks
QUICK SEARCH Keyword SEARCH ADVANCED SEARCH
FREE Shipping When you buy two or more items!
See Details
Related Information
Bibliography
Information Rules: A Strategic Guide to the Network Economy
Carl Shapiro Carol Shapiro Hal R. Varian
Retail Price: $29.95
Our Price: $23.96
You Save: $5.99 (20%)
Readers' Advantage Price:
$22.76 Join Now
In Stock:Ships within 24 hours
Same Day Delivery in Manhattan.
Format: Hardcover, 368pp.
ISBN: 087584863X
Publisher: Harvard Business School Publishing
Pub. Date: November 1998
Buy it Now!
Add To
As you order, each item will be listed in Your Shopping Cart in the upper right corner. You may make changes at Checkout.
Safe Shopping Guarantee!
Done
Internet

8



Predicting Purchase Conversion

Home

Category

Product

Shopping Cart

What is the chance of this user making a purchase during this session?

- 1st viewing = 7%
- 2nd viewing = 14%
- 3rd viewing = 20%
- 4th viewing = 60%

Problems with Clickstream Data

- Huge
  - Terabytes of data
- Collection Issues
  - Server versus client
  - Privacy and identifiability
- Unstructured
  - Textual
- Decision Problem
  - Redesign web site
  - Understand usage and implication on consumer behavior

Data

Data

- Comscore Media Metrix Panelists from April 2002 who visit barnesandnoble.com
  - 1,160 users
  - 1,659 Sessions, 8.75 Viewings per session
  - 14,512 viewings
  - 114 purchases ⇒ 7% conversion rate
- Collect HTML content using Perl script
  - Parse the pages for content information, like the presence of a banner ad, price, …

Path analysis can be done using server logs from bn.com directly

Page Category Examples

Home

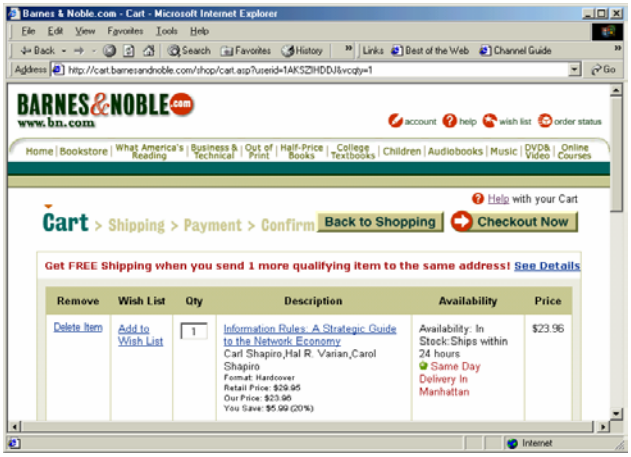
Category

Product

Information

Account

Shop. Cart



Categorization of Pages

Abbr-eviation	Category Name	Description	% of Page Requests
H	Home	Home page, common starting page for B&N visitors	6.5%
A	Account	User account pages that allow a user to sign in, change address, and review order status	6.4%
C	Category	Pages that have list of items (links like cooking, fiction, etc. from the home page) or results of a book search	32.9%
P	Product	Pages with detailed product information, item description, price information, availability, and reviews	16.4%
I	Information	Pages with shipping, order status, and popup advertisements such as “Free Shipping”	29.1%
S	Shopping Cart	Pages related to the shopping cart, such as reviewing the cart, deleting items, entering purchase information	7.0%
O	Order	Confirmation page that denotes order has taken place	1.0%
E	Enter/Exit	Non B&N pages used to denote the beginning or end of a session	.8%

Table 3. Categorization scheme for B&N web pages for 9,180 unique pages requests.

Transformed Clickstream

	Time	URL	Category	Abbr.
1	8:36:11pm	/promo/coupon/popups/fs_usa_popup.asp?userid=xxx	Information	I
2	8:36:29pm	/booksearch/results.asp?wrd=70%2d215&userid=xxx	Category	C
3	8:36:48pm	/booksearch/results.asp?userid=xxx&mscssid=yyy&wrd=70%2d215&opr=a&sort=p	Category	C
4	8:37:14pm	/booksearch/isbninquiry.asp?userid=xxx&mscssid=yyy&isbn=0072134445	Product	P
5	8:38:10pm	/booksearch/results.asp?userid=xxx&mscssid=yyy&wrd=70%2d215&opr=a&sort=p	Category	C
6	8:44:32pm	/textbooks/booksearch/isbninquiry.asp?userid=xxx&mscssid=yyy&isbn=0619034971	Product	P
7	8:55:12pm	/promo/coupon/popups/fs_usa_popup.asp?userid=xxx	Information	I
8	8:55:24pm	/booksearch/results.asp?wrd=70%2d215&userid=xxx	Category	C
9	8:55:36pm	/booksearch/results.asp?userid=xxx&mscssid=yyy&wrd=70%2d215&opr=a&sort=p	Category	C
10	8:56:37pm	/shop/signin.asp?userid=xxx&mscssid=yyy	Account	A
11	8:58:16pm	/booksearch/results.asp?userid=xxx&mscssid=yyy&wrd=70%2d215&opr=a&sort=p	Category	C
12	8:58:40pm	/booksearch/isbninquiry.asp?userid=xxx&mscssid=yyy&isbn=0072224983	Product	P
13	8:59:21pm	/shop/cart.asp?userid=xxx&mscssid=yyy	Account	C
14	9:01:26pm	Exit	Exit	E

Some Sample User Sessions

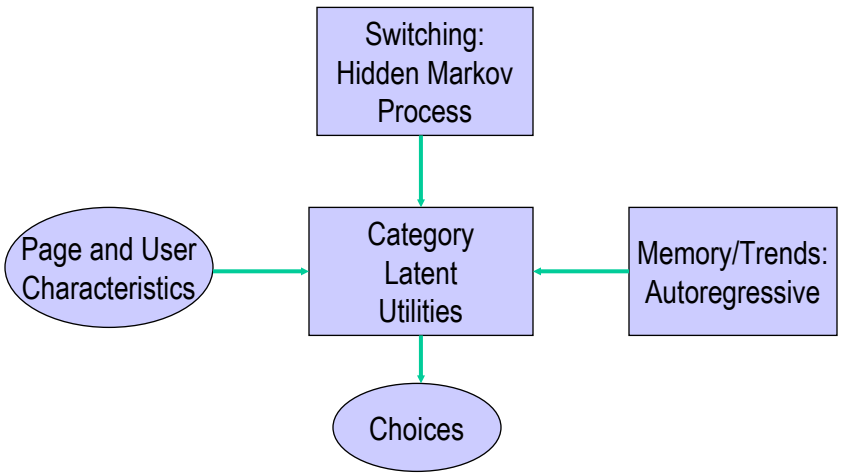
	User	Path
Browsers	1	ICCCCCCCCCPCCPCCCCCCCCCCCCCCCCCCCCCCCCCCCCE
	2	IHHE
	3	IE
	4	IHICPPPCE
	5	IHHIIICIIE
Buyers	6	HIAAAAIAIIIICIIICICICCICICCIPPIPPIPPPIPIICCSIIIPPPPPIPIPSISISISSSOIIIIIHE
	7	HCCPPPCCPCCCCCCCCCPSCSCSPCCPCPCCCCCCCSAAAAAAAAAASSOIIIISASCCCE
	8	IIICICPCPPPCPCICICPCCCPCPPPIPSIIAASSSIIIIOIIE
	9	IISIASSSIOIE
	10	IPPPPSASSSSSOIAAAHCCPCCCCCCE

Model

Model Elements

- Dependent Variable is a Categorical Time Series
  - Choice modeled using Multinomial Probit (MNP Model)
- Two Types of Time Series Behavior:
  - Continuous trends: VAR(1) Process
  - Abrupt changes: Mixture Process with a Hidden Markov Model
- Covariates
  - Web design and marketing mix
  - Browsing behavior
  - Demographics
- Consumer Heterogeneity
  - Hierarchical Bayesian Approach

Choice Model of Browsing



Dynamic Multinomial Probit Model with A Mixture Process

- Continuous-Time Hidden Markov Chain D_i with λ_s, the Waiting Time Intensity Parameter

$$\mathbf{P}_i = \begin{bmatrix} 0 & P_{12i} & \cdots & P_{1Si} \\ P_{21i} & 0 & \cdots & P_{2Si} \\ \vdots & \vdots & \ddots & \vdots \\ P_{S1i} & P_{S2i} & \cdots & 0 \end{bmatrix}, \quad \mathbf{v}_i = \begin{bmatrix} v_{1i} \\ v_{2i} \\ \vdots \\ v_{Si} \end{bmatrix}$$

- "Time" can refer to either per viewing, per session, or per user (which is a latent mixture model)

Dynamic Multinomial Probit Model with A Mixture Process

- Observation Equation:

$$Y_{ict} \begin{cases} = 1 \quad if \quad \max(U_{ikt}) < 0 \\ = c \quad if \quad U_{ict} \geq \max(U_{ikt}) \end{cases}$$

- Model Setup:

Memory | Covariates | Hidden Markov States

$$U_{ict} = \boldsymbol{\Phi}_{ics} \boldsymbol{U}_{i(t-1)} + \boldsymbol{\gamma}_{ics}' \boldsymbol{X}_{it} + \varepsilon_{icts}$$

Latent Utility

with probability p_s such that $\sum_{s=1}^{S} p_s = 1$

Understanding State Model

- Continuous Time Model

$$f(D_i \mid P_i, \lambda_i, \nu_i) = (\prod_s \nu_i^{\mathrm{I}(D_{i0}=s)})(\prod_{s \neq l} P_{isl}^{N_{isl}})(\prod_s \lambda_{is}^{M_{is}}) \exp[-\sum_s (\int_0^T \lambda_{is}\, \mathrm{I}(D_{it}=s)dt)]$$

- Per viewing approximates First-Order Markov Model

$$\mathbf{P} = \begin{bmatrix} P_{11} & P_{12} \\ P_{21} & P_{22} \end{bmatrix} = \begin{bmatrix} 1-\exp\{-\lambda_1\} & \exp\{-\lambda_1\} \\ \exp\{-\lambda_2\} & 1-\exp\{-\lambda_2\} \end{bmatrix}$$

Other Model Elements

- Hierarchical Bayesian Model
  - Parameters are "regressed" upon user characteristics
- Model includes many special cases:
  - Multinomial Probit Model (1 state)
  - Latent Mixture Model (Multiple states, zero-order markov)
  - Zero-Order Markov Process (only intercepts)
  - Approximates First-Order Markov Process with VAR(1)

Predictive Results

Page Transition Models

State Time	State Process	Number of States	Log Marginal Density	In-Sample Hit Rate (%)	Out-of-Sample Hit Rate (%)
Page	Zero-Order	1	-9378.1	72.05	65.40
		2	-9016.9	79.44	71.42
		3	-9064.0	80.34	70.56
	First-Order	1	-8545.4	83.23	79.95
		2	-8428.3	89.71	83.15
		3	-8474.0	89.97	81.14

Best

Session Transition Models

State Time	State Process	Number of States	Log Marginal Density	In-Sample Hit Rate (%)	Out-of-Sample Hit Rate (%)
Session	Zero-Order	1	-9376.1	73.17	61.56
		2	-9051.0	77.90	70.48
		3	-9097.7	78.76	66.14
	First-Order	1	-8573.5	83.05	73.57
		2	-8464.9	88.44	81.48
		3	-8487.0	88.73	78.42

User Transition Models

State Time	State Process	Number of States	Log Marginal Density	In-Sample Hit Rate (%)	Out-of-Sample Hit Rate (%)
User	Zero-Order	1	-9411.1	64.38	61.50
		2	-9124.2	70.04	64.12
		3	-9193.8	70.85	63.99

Table 7. Measures of fit for various dynamic probit models.

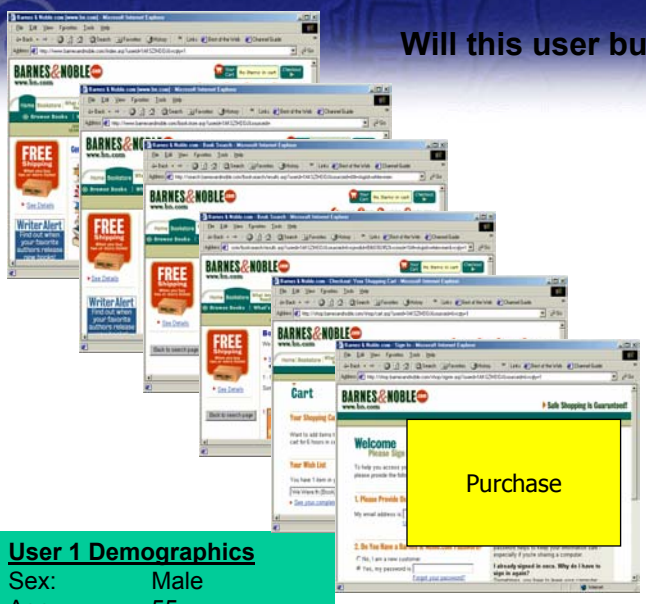
Alternative Model Specifications

Model	Log Marginal Density	In-Sample Hit Rate (%)	Out-of-Sample (%)
Zero-Order Markov Model (1 State)	-20410.4	20.48	12.62
Zero-Order Markov Model (2 States)	-19458.3	28.18	19.02
First-Order Markov Model (1 State)	-16444.5	56.06	51.59
First-Order Markov Model (2 States)	-16076.0	58.61	52.08
Latent Class Model (1 State)	-17849.2	35.47	30.78
Latent Class Model (2 States)	-17673.9	44.29	40.21
Latent Class Model (3 States)	-17722.3	45.29	36.14
Independent	-19086.4	33.23	30.35
Only-Intercept	-19335.9	29.37	23.12
Intercept + VAR	-13768.4	71.13	64.38

Summary of Findings

- Banner ads tend to encourage browsing oriented individuals but discourage continued browsing from deliberation oriented users
- Including price information tends to discourage browsing oriented users but encourage deliberation oriented users
- More home links leads to less use of home page by browsers and more by deliberation oriented users
- Weekend users more likely to surf
- Browsers who have purchased before are more likely to order
- Rich set of relationships in dynamic browsing behavior

Purchase Conversion



Will this user buy?

13.8%

12.3%

13.2%

14.3%

35.3%

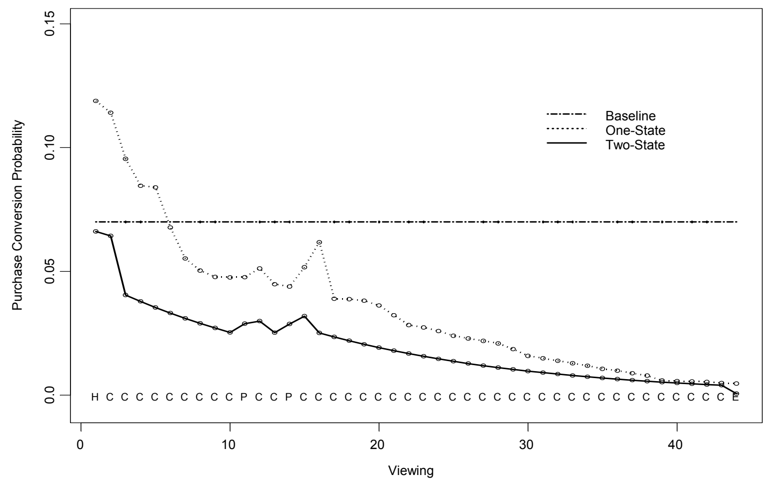
52.4%

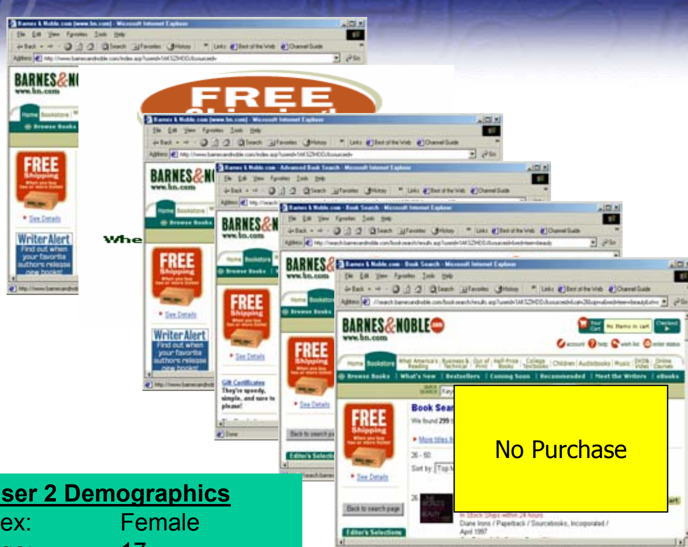
.
.
.

User 1 Demographics

Sex:	Male
Age:	55
Occupation:	Service Worker
State:	Washington

User 1: Purchase Conversion Predictions



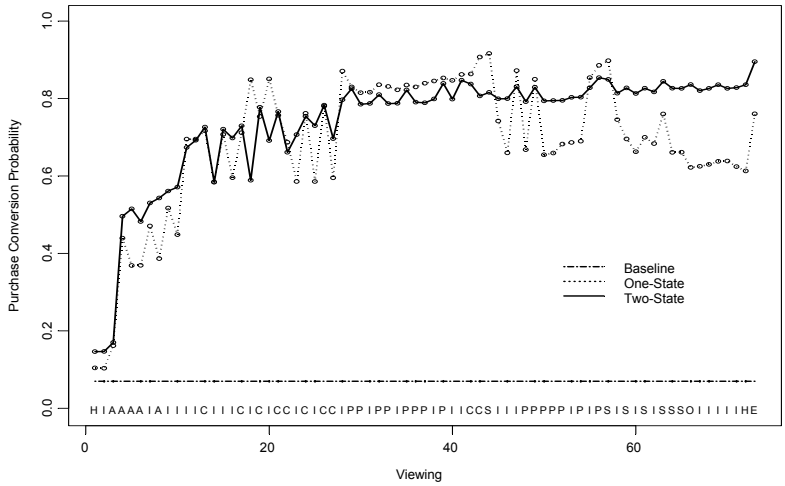


0.24%
0.26%
0.06%
0.05%
0.04%
0.03%
.
.
.

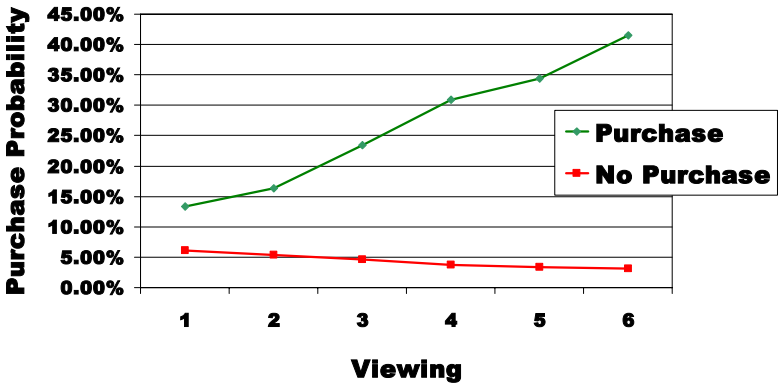
User 2 Demographics

Sex: Female
Age: 17
Occupation: Student
State: Virginia

User 2: Purchase Conversion Predictions



Predicting Purchase Conversion



Predicted Purchase Conversions

Sample	Session Type	Number of Sessions	Forecast Origin/Number of viewings during session					
			1	2	3	4	5	6
Estimation	Purchase	83	13.3% (0.48)	16.3% (0.52)	23.4% (0.60)	30.9% (0.65)	34.4% (0.67)	41.5% (0.69)
	No Purchase	1129	6.1% (0.33)	5.4% (0.32)	4.6% (0.30)	3.7% (0.27)	3.4% (0.26)	3.1% (0.25)
	All	1212	6.6% (0.35)	6.1% (0.34)	5.9% (0.33)	5.6% (0.33)	5.5% (0.32)	5.7% (0.33)
Holdout	Purchase	31	10.4% (0.97)	12.8% (1.06)	15.2% (1.14)	18.0% (1.21)	19.1% (1.24)	21.2% (1.29)
	No Purchase	416	6.9% (0.80)	5.5% (0.72)	5.1% (0.70)	4.2% (0.63)	3.5% (0.58)	3.2% (0.56)
	All	447	7.2% (0.82)	5.9% (0.75)	5.8% (0.74)	5.1% (0.70)	4.6% (0.66)	4.4% (0.65)

Conclusions

Conclusions and Managerial Implications

- Possible to incorporate pathing information on user browsing
- Good dynamic prediction of purchase conversion
- Measurable impact of marketing mix and hypertext links

Limitations and future directions

- One retailer for an one-month period
- Categorization of the web pages
  - Predict choices at a link level instead of category
- Structural model