Personalization from Incomplete Data:
What You Don't Know Can Hurt

Balaji Padmanabhan
Zhiqiang Zheng
Steven O. Kimbrough

Presented by Zhenchang Xing

Outline

- **Introduction**
- The Methodology
  - Data and Usage Metrics
  - Classification Models and Evaluation Criteria
- Results
- Conclusion

Introduction

- Personalization
  - In industry
  - In academia
- The problem?
  - Built on data collected by *a single web site*

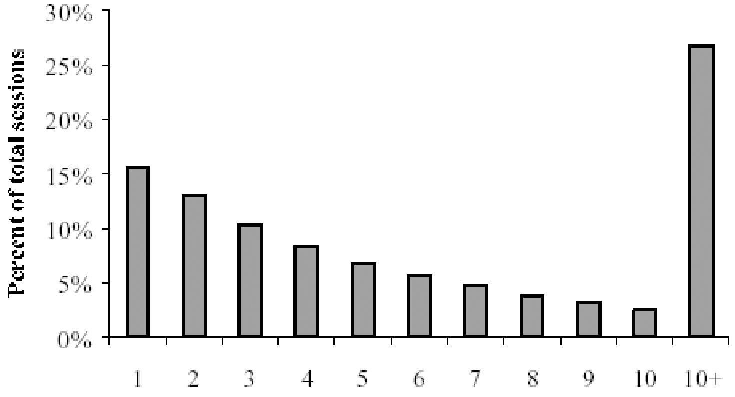
Example Sessions

- User 1: Cheaptickets_1, Cheaptickets_2, Travelocity_1, Travelocity_2, Expedia_1, Expedia_2, Travelocity_3, Travelocity_4, Expedia_3, Cheaptickets_3
  - Assume that this user bought a ticket at Cheaptickets
- User 2: Expedia_1, Expedia_2, Expedia_3, $\mathbf{Expedia_4}$
  - Assume that this user bought at Expedia_4

Expedia Sees

- User1: Expedia$_1$, Expedia$_2$, Expedia$_3$
  - No buying
- User2: Expedia$_1$, Expedia$_2$, Expedia$_3$, Expedia$_4$
  - Buying at Expedia

Sites Visited in a Session



Definitions

- Site-Centric Data
  - web log + user demographics
- User-Centric Data
  - ‘Complete’ version of usage data, Purely hypothetical
- Session-Level Prediction
  - Whether the remainder of a current user’s session will result in a purchase
- User-Level Prediction
  - Whether a given user at a given point in time will make a purchase at the site during some future session

Outline

- Introduction
- **The Methodology**
  - Data and Usage Metrics
  - Classification Models and Evaluation Criteria
- Results
- Conclusion

The Methodology

- Starting with raw data provided by MediaMetrix
- Construct site-centric and user-centric data from raw user-level browsing data
- Preprocessing for two-level prediction tasks
  - Preprocessing for Session-Level Prediction
  - Preprocessing for User-level Prediction
- Build 4 different classifiers for two-level predictions based on two types of preprocessed datasets (40% training set, 60% evaluation)
- Compare the performance of 8 pairs of classification models quantitatively and qualitatively

Raw Data

- Raw data provided by Media Metrix
  - 20,000 user's web browsing behavior over 6 months
  - 30GB and 4 million user sessioins
  - User demographics
  - Transaction history over the entire period
  - Sites categories: book, music, travel, auction, general shopping mall (310,323 user sessions, 135 web sites)
- The tracking software installed on the client machine

Construct site-centric data and user-centric data

Cheaptickets$_1$, Cheaptickets$_2$, Travelocity$_1$, Travelocity$_2$, Expedia$_1$, Expedia$_2$, Travelocity$_3$, Travelocity$_4$, Expedia$_3$, Cheaptickets$_3$

- Site-Centric Data
  - Cheaptickets$_1$, Cheaptickets$_2$, Cheaptickets$_3$
  - Travelocity$_1$, Travelocity$_2$, Travelocity$_3$, Travelocity$_4$
  - Expedia$_1$, Expedia$_2$, Expedia$_3$
- User-Centric Data

Usage Metrics

- Current visit summaries, e.g. time spent in current session
- Historical summaries, e.g. average time spent per session in the past
- User demographics, e.g. name, gender

Site-Centric Data Preprocessing for Session-Level Prediction

- Consider a single session of length 5
  - <p_1, p_2, p_3, $\boldsymbol{p_4}$, p_5>.
- This single sessions generates 5 records for prediction
  - 1. A session that began with p_1 resulted in the user booking at a subsequent point.
  - 2. A session that began with p_1, p_2 resulted in booking at a subsequent point.
  - 3. A session that began with p_1, p_2, p_3 resulted in booking at a subsequent point.
  - 4. A session that began with p_1, p_2, p_3, p_4 did *not* result in booking at a subsequent point.
  - 5. A session that began with p_1, p_2, p_3, p_4, p_5 did *not* result in booking at a subsequent point.

Site-Centric Data Preprocessing for Session-Level Prediction-cont’

- Probabilistic Sampling
  - A session of length k on average provides α*k records
- Probabilistic Clipping
  - Every sampled session is clipped probabilistically based on its length and divided into two parts
    - The first part will be used to compute usage metrics
    - The second part will be used to determine whether a purchase occurred
      - Heuristic, such as user time spent under secure-mode
- Usage Metrics
  - 6 demographic + 5 Historical + 4 Current + 1 Site Category

Site-Centric Data Preprocessing for Session-Level Prediction-cont’

- Consider a single session of length 5
  - <p_1, p_2, p_3, $\boldsymbol{p_4}$, p_5>.
- Sample rate = 0.4 Clipping point 1 and 3
  - 1. A session that began with p_1 resulted in the user booking at a subsequent point.
  - 3. A session that began with p_1, p_2, p_3 resulted in booking at a subsequent point.

User-centric data preprocessing for Session-Level Prediction

- Probabilistic Clipping augmented with what else sites the user visited
  - User session: C_1, C_2, T_1, T_2, E_1, E_2, T3, T4, E_3, C_3
  - Site-Centric data for site E: E_1, E_2, E_3
  - User-Centric data for site E: C_1, C_2, T_1, T_2, E_1, E_2, T3, T4, E_3, C_3
  - Clipping point: E_1
  - User-Centric clipping point: C_1, C_2, T_1, T_2, E_1
- Usage metrics
  - 17 Historical + 8 Current additional metrics

Data Preprocessing for User-level Prediction

- For user U and web site E
  - N user sessions in raw data involving E
    - C_1, C_2, T_1, T_2, E_1, E_2, T3, T4, E_3, C_3
  - N site-centric sessions for E: s_1, s_2, … s_n
    - E_1, E_2, E_3
  - N user-centric sessions for E: u_1, u_2, … u_n
    - C_1, C_2, T_1, T_2, E_1, E_2, T3, T4, E_3, C_3
  - N summary records at the end of each sesssion
- Usage Metrics
  - Site-Centric sessions: 6 demographic + 5 Historical + 1 Site Category
  - User-centric sessions: 6 demographic + 17 Historical + 1 Site Category

Outline

- Introduction
- The Methodology
  - Data and Usage Metrics
  - **Classification Models and Evaluation Criteria**
- Results
- Conclusion

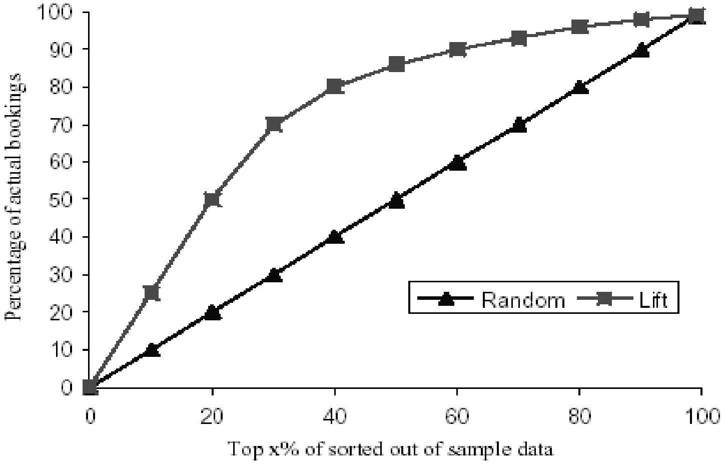
Classification Models

- 4 Classification Models
  - Liner regression (linear)
  - Logistic regression (log-linear)
  - Classification tree (non-linear)
  - Neural network (non-linear)
- The reason to choose
  - Data-driven
  - Linear, log-linear and non-linear

Evaluation Criteria

- Quantitative Comparison
  - Prediction accuracy
    - Limitation: unequal priors
  - Lift curves
    - Binary prediction
    - Classification models provide a kind of probability or confidence measure in the predicted value
- Qualitative Insights Analysis
  - Consistency
  - Contradiction
  - Incompleteness

An Example of Lift Curves



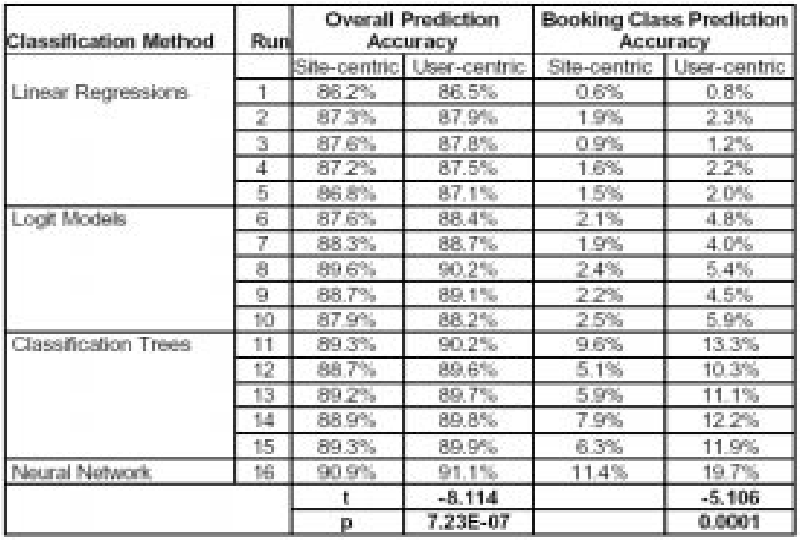
The Methodology

- Starting with raw data provided by MediaMetrix
- Construct site-centric and user-centric data from raw user-level browsing data
- Preprocessing for two-level prediction tasks
  - Preprocessing for Session-Level Prediction
  - Preprocessing for User-level Prediction
- Build 4 different classifiers for two-level prediction based on two types of preprocessed datasets (40% training set, 60% evaluation)
- Compare the performance of 8 pairs of classification models quantitatively and qualitatively

Outline

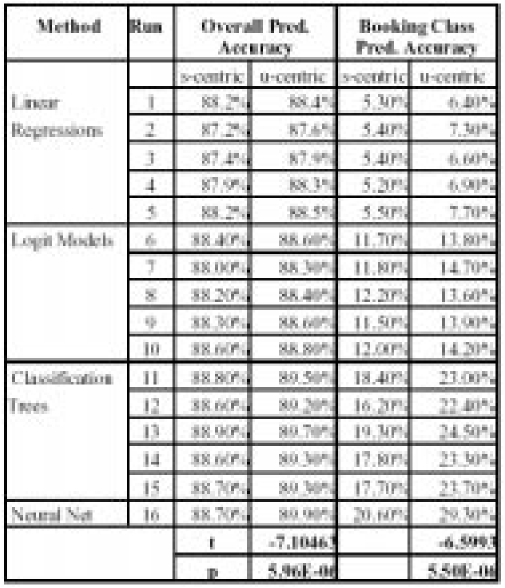
- Introduction
- The Methodology
  - Data and Usage Metrics
  - Classification Models and Evaluation Criteria
- **Results**
- Conclusion

Prediction Accuracy: Site-Level Prediction



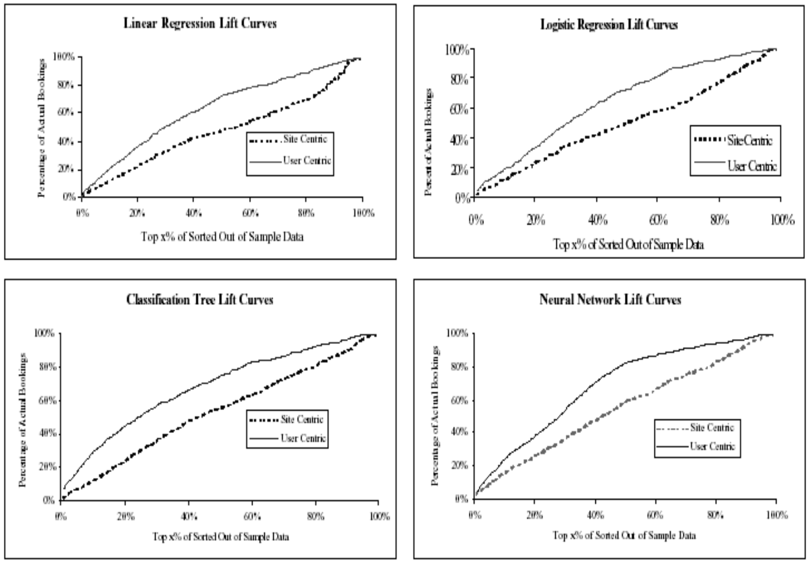
Classification Method	Run	Overall Prediction Accuracy		Booking Class Prediction Accuracy	
		Site-centric	User-centric	Site-centric	User-centric
Linear Regressions	1	86.2%	86.5%	0.6%	0.8%
	2	87.3%	87.9%	1.9%	2.3%
	3	87.6%	87.8%	0.9%	1.2%
	4	87.2%	87.5%	1.6%	2.2%
	5	86.8%	87.1%	1.5%	2.0%
Logit Models	6	87.6%	88.4%	2.1%	4.8%
	7	88.3%	88.7%	1.9%	4.0%
	8	89.6%	90.2%	2.4%	5.4%
	9	88.7%	89.1%	2.2%	4.5%
	10	87.9%	88.2%	2.5%	5.9%
Classification Trees	11	89.3%	90.2%	9.6%	13.3%
	12	88.7%	89.6%	5.1%	10.3%
	13	89.2%	89.7%	5.9%	11.1%
	14	88.9%	89.8%	7.9%	12.2%
	15	89.3%	89.9%	6.3%	11.9%
Neural Network	16	90.9%	91.1%	11.4%	19.7%
		t	**-8.114**		**-5.106**
		p	**7.23E-07**		**0.0001**

Prediction Accuracy: User-Level Prediction

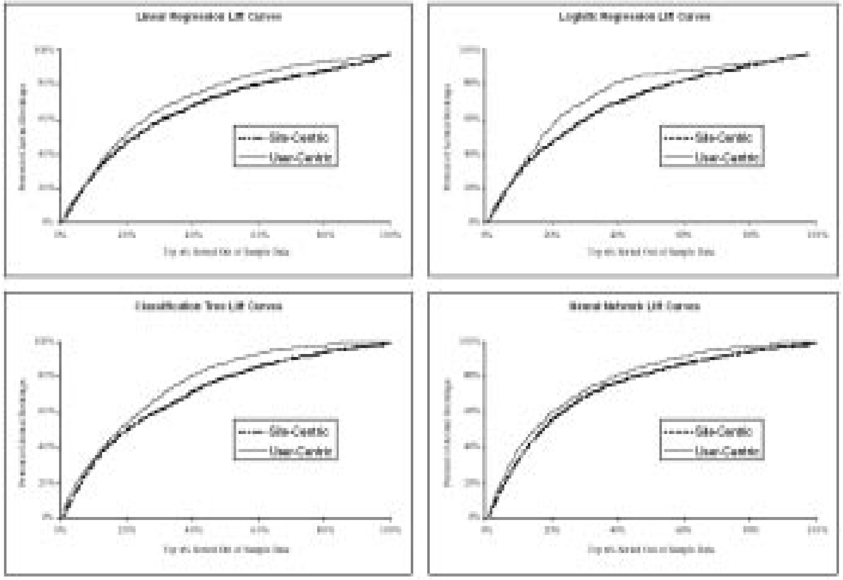


Method	Run	Overall Pred. Accuracy		Booking Class Pred. Accuracy	
		s-centric	u-centric	s-centric	u-centric
Linear Regressions	1	88.2%	88.4%	5.30%	6.40%
	2	87.2%	87.6%	5.40%	7.30%
	3	87.4%	87.9%	5.40%	6.60%
	4	87.9%	88.3%	5.20%	6.90%
	5	88.2%	88.5%	5.50%	7.70%
Logit Models	6	88.40%	88.60%	11.70%	13.80%
	7	88.00%	88.30%	11.80%	14.70%
	8	88.20%	88.40%	12.20%	13.60%
	9	88.30%	88.60%	11.50%	17.90%
	10	88.60%	88.80%	12.00%	14.20%
Classification Trees	11	88.80%	89.50%	18.40%	23.00%
	12	88.60%	89.20%	16.20%	22.40%
	13	88.90%	89.70%	19.30%	24.50%
	14	88.60%	89.30%	17.80%	23.30%
	15	88.70%	89.30%	17.70%	23.70%
Neural Net	16	88.70%	89.90%	20.60%	29.30%
		t	-7.10463		-6.5993
		p	5.96E-06		5.50E-06

Lift Curves: Session-Level Prediction



Lift Curves: User-Level Prediction



Qualitative Insight Analysis

- Consistency
  - Purchase in the past highly positively correlated with potential current session purchase
- Contradiction
  - Using site-centric data, total time spent at a current session is highly important, but this effect does not hold for user-centric data
- Incompleteness
  - Purchase at *any* site are very significant across all models

Conclusion

- Models built from complete data (user-centric) significantly outperform ones derived from incomplete data (site-centric)
- Potentially erroneous conclusions can be drawn from incomplete data
- The effects may vary based on different tasks considered