电子商务数据分析

第6章 在线购买预测

朱桂祥 (9120201070@nufe.edu.cn)
南京财经大学信息工程学院
江苏省电子商务重点实验室
电子商务信息处理国家级国际联合研究中心
电子商务交易技术国家地方联合工程实验室

湖南大学大学校训

实事求是
敢为人先

“实事求是”指立足现实，夯实基础，追求真理，脚踏实地；**“敢为人先”**指着眼于未来和长远，敢于竞争、敢于创新，走特色发展的道路。**“实事求是”**是对岳麓书院“通经致用”“爱国务实”“重践履、务实学、通时务”精神的发展，**“敢为人先”**则是对以岳麓书院为代表的湖湘学派以及湖湘文化精髓“敢为天下先”精神的传承

https://tv.cctv.com/2014/08/11/VIDE1407715503559422.shtml

目 录
Contents

01 购买预测简介

一、购买预测简介

1. 背景

- ✓ 随着越来越多的电子商务平台的兴起(如亚马逊,淘宝,京东等),能刻画用户在线行为和兴趣偏好的在线数据越来越丰富.

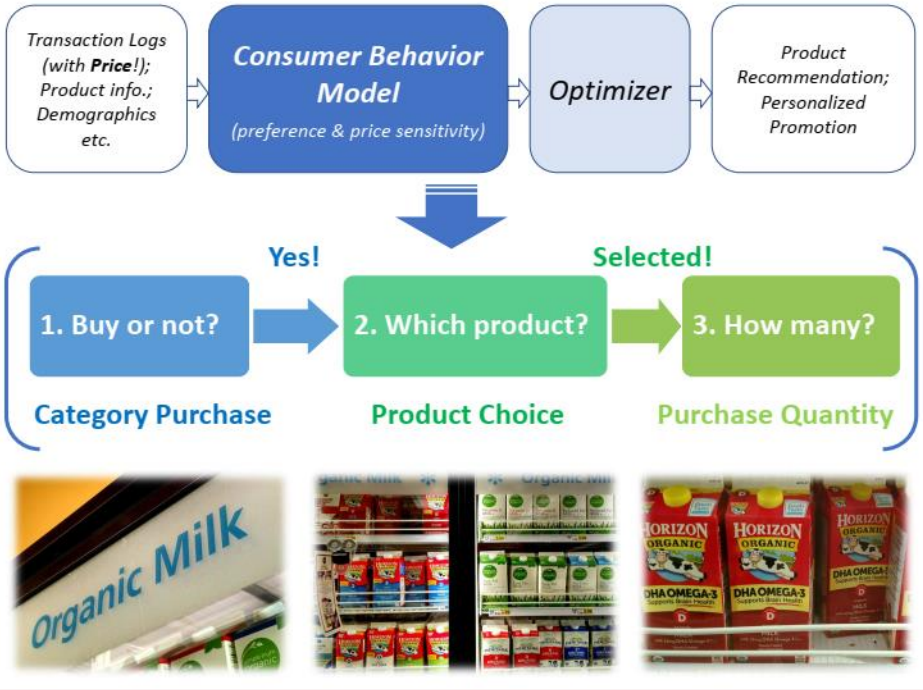
- ✓ 企业亟需借助大数据分析技术来提升用户智能化服务,整合市场营销及运营管理,增加经济效益,这使得在线购买决策成为商务智能研究领域的热门议题之一.

- ✓ 企业需要在海量数据挖掘基础上建立一种高级商务智能平台,以帮助电子商务网站为其顾客购物提供个性化的决策支持和信息推荐服务.

一、购买预测简介

1. 背景

- 经典的购买决策模型通常细化为3个阶段：
  1. 是否买?
  2. 买什么?
  3. 买多少?



Search

Recommendation

亚马逊 amazon.cn

淘

京东

一、购买预测简介

1. 背景

工业界竞赛 (Challenge):

阿里云 TIANCHI天池

✓ 阿里巴巴天池大数据竞赛

The ACM Conference Series on Recommender Systems

✓ **Recsys 2015 Challenge**

【The Task】

Given a sequence of click events performed by some user during a typical session in an e-commerce website, the goal is to predict whether the user is going to buy something or not, and if he is buying, what would be the items he is going to buy. The task could therefore be divided into two sub goals:

1.Is the user going to buy items in this session? Yes|No

2.If yes, what are the items that are going to be bought?

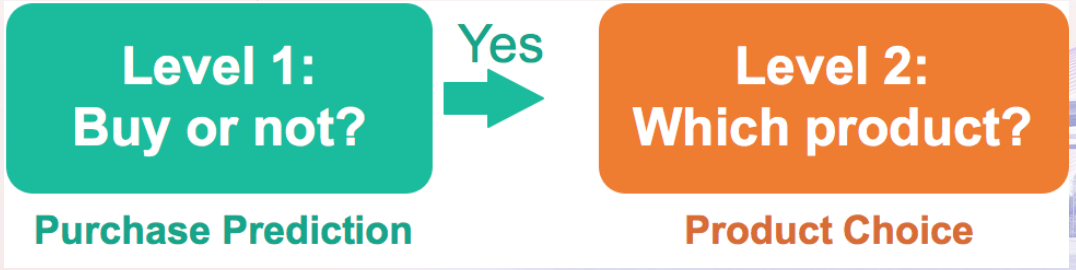
一、购买预测简介

2. 在线旅游购买决策

在旅游电子商务中，越来越多的游客通过各种各样的在线平台收集更丰富，全面，个性化的旅游信息用于他们的旅游行程规划. 因此，产生了海量的在线旅游数据，电子商务旅游平台也亟需通过一些新颖的数据分析和挖掘的技术手段去实现商务的潜能.

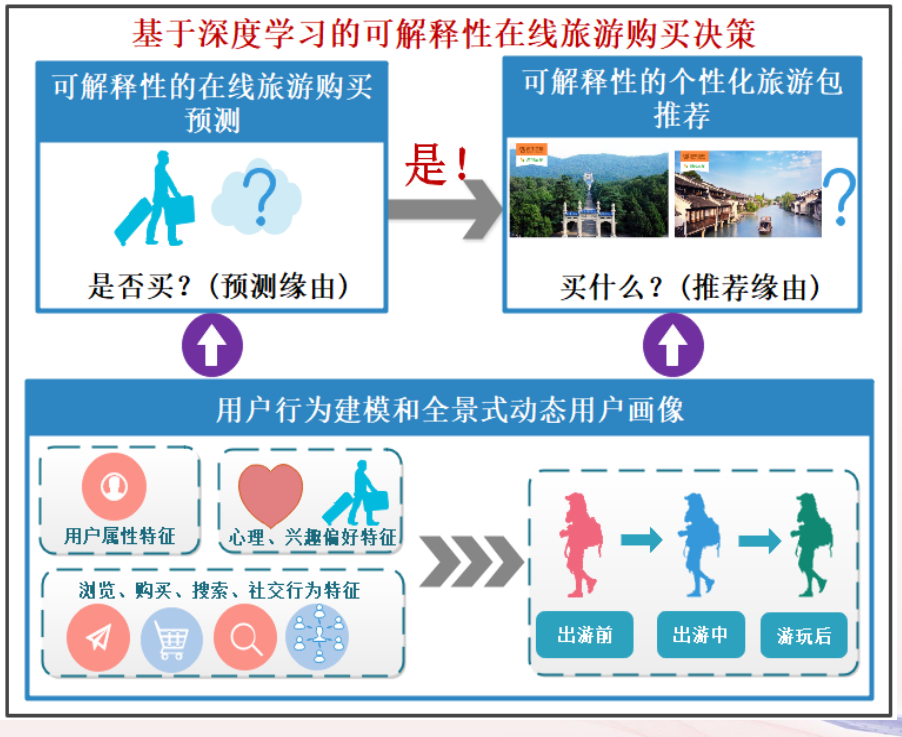
Expedia® TRAVELZOO® 旅游族 马蜂窝 途牛旅游网 Tuniu.com

考虑到旅游产品通常不存在购买数量上的差异，我们拟沿着两阶段购买决策模型的宏观框架对电子商务旅游数据展开具体方法的研究. 希望在线旅游购买决策模型能够回答以下2个层面的问题：



一、购买预测简介

2. 在线旅游购买决策



从企业角度来看，本项目的成功实施预期能够提升旅游企业自身影响力和竞争力，有效树立企业品牌形象，给旅游业带来新的发展动力，大大降低交易成本和生产管理成本，为企业创造更多的利润，促进旅游信息化发展，提升旅游企业个性化服务的能力。从行业角度来看，本项目的顺利实施能够加快大数据技术在旅游业的应用、普及与提高，积极探索旅游电子商务的新模式、培育新业态，对我国旅游业及相关企业信息化建设起到示范作用。

1. Zhu G, Wu Z, Wang Y, et al. Online purchase decisions for tourism e-commerce. *Electronic Commerce Research and Applications*, 2019, 38: 100887.

02 在线购买预测建模

二、在线购买预测建模

途牛旅游网 Tuniu.com

2. 相关工作：

	标题	期刊/会议	作者
购买分析	A model of web site browsing behavior estimated on clickstream data	JMR 2003	Bucklin等人
	Modeling purchase behavior at an e-commerce web site: A task-completion approach	JMR 2004	Sismeiro等人
	Converting web site visitors into buyers: How web site investment increases consumer trusting beliefs and online purchase intentions	JM 2006	Schlosser等人
	Estimating product-choice probabilities from recency and frequency of page views	KBS 2016	Iwanaga等人
	An extended online purchase intention model for middle-aged online users	ECRA 2016	Law等人
购买预测	Predicting purchase behaviors from social media	WWW 2013	Zhang等人
	An ensemble approach for multi-label classification of item click sequences	Recsys 2015 Challenge	Yağci等人
	Repeat buyer prediction for e-commerce	SIGKDD 2016	Liu等人
	Predicting shopping behavior with mixture of RNNs	SIGIR 2017 Workshop	Toth等人

相关工作总结

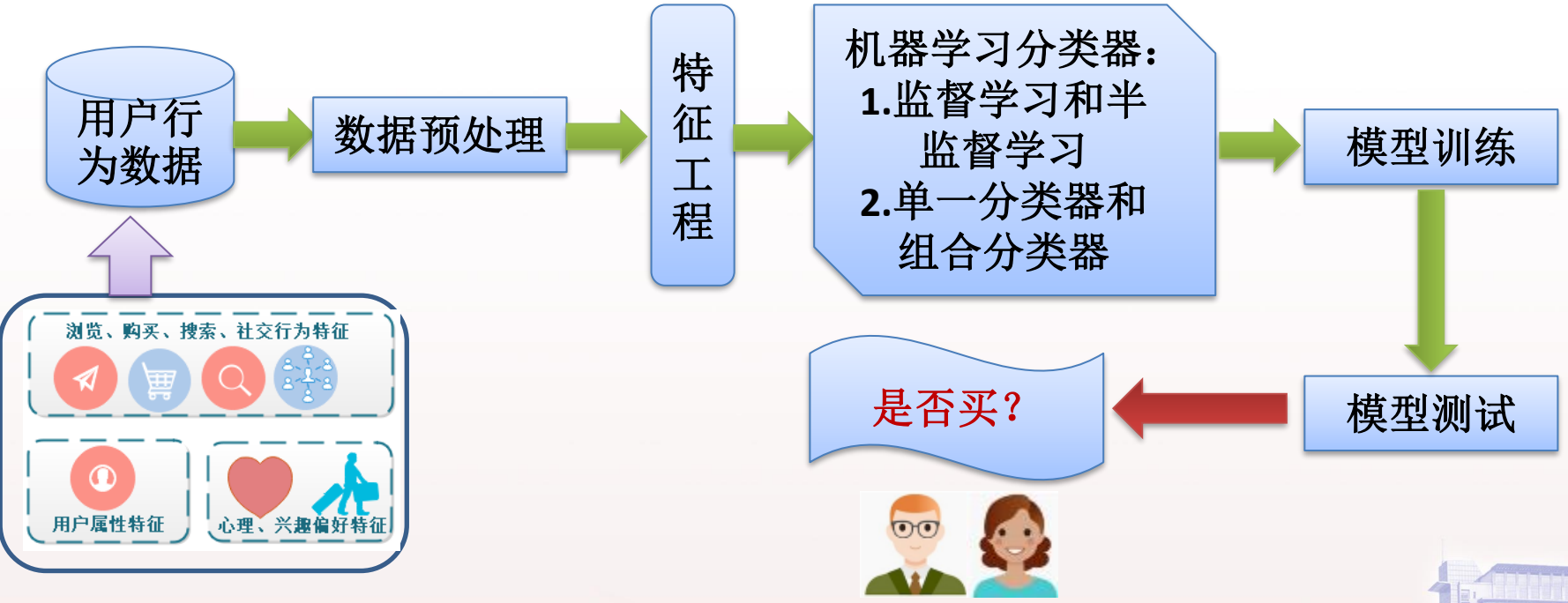
此类研究基本建立了访问和购买之间的一般性因果关系,并揭示出购买行为是经过深思熟虑后的策划结果,即所谓的计划行为理论.

此类研究致力于建立因果型购买决策模型,即回答有购买意向的客户具备哪些特征(或受哪些因素影响).

另一类研究则试图建立预测型购买决策模型,即预测哪些客户有(无)购买意向.已有工作主要集中在特征工程的构造上,而分类模型多直接使用经典的决策树或者组合分类器等.

二、在线购买预测建模

■ 机器学习方法：

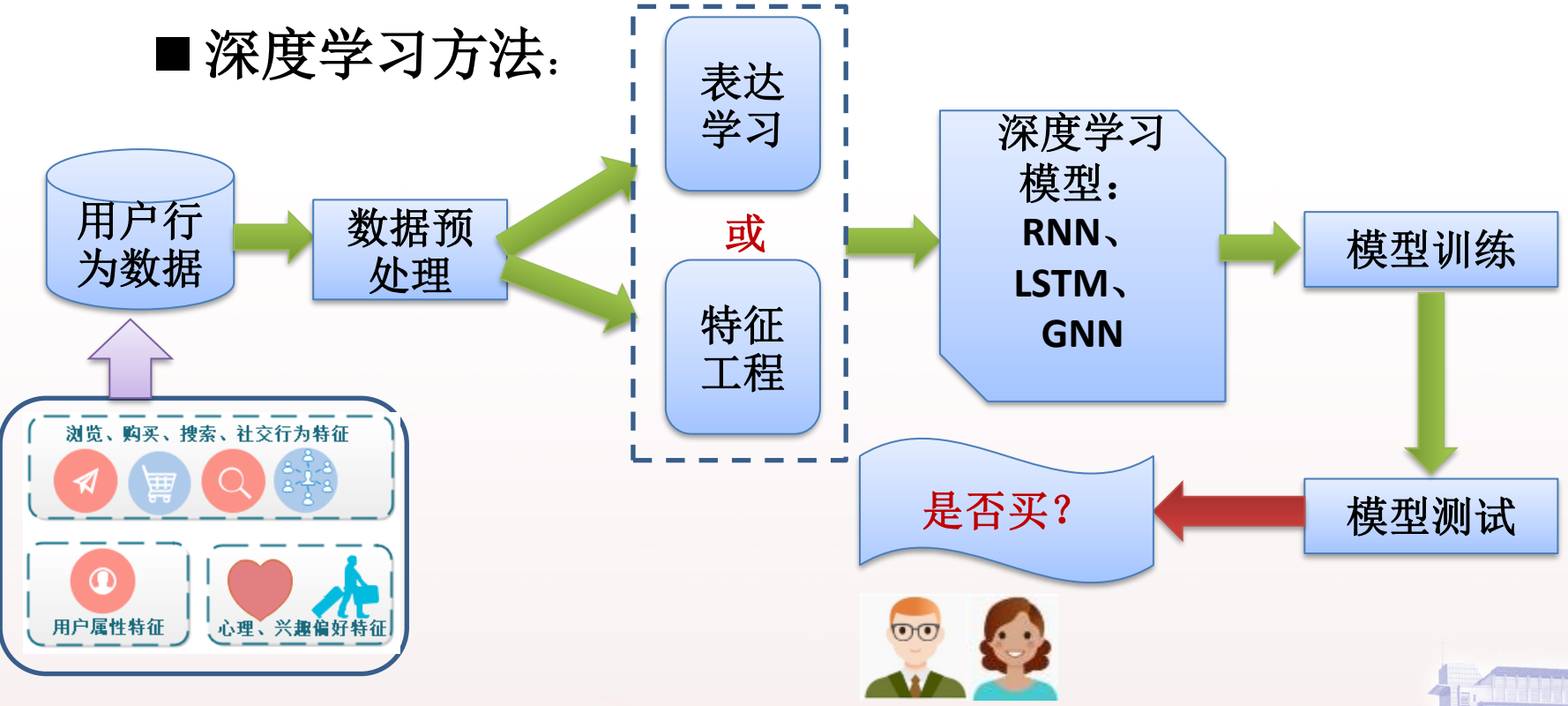


[1]Zhu G, Wu Z, Wang Y, et al. Online purchase decisions for tourism e-commerce[J]. Electronic Commerce Research and Applications, 2019, 38: 100887.

[2]Li D, Zhao G, Wang Z, et al. A method of purchase prediction based on user behavior log[C]//2015 IEEE International Conference on Data Mining Workshop (ICDMW). IEEE, 2015: 1031-1039.

二、在线购买预测建模

■ 深度学习方法：



[1] Ling C, Zhang T, Chen Y. Customer purchase intent prediction under online multi-channel promotion: A feature-combined deep learning framework[J]. IEEE Access, 2019, 7: 112963-112976.

[2] Guo L, Hua L, Jia R, et al. Buying or browsing?: Predicting real-time purchasing intent using attention-based deep network with multiple behavior[C]//Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 2019: 1984-1992.

二、在线购买预测建模

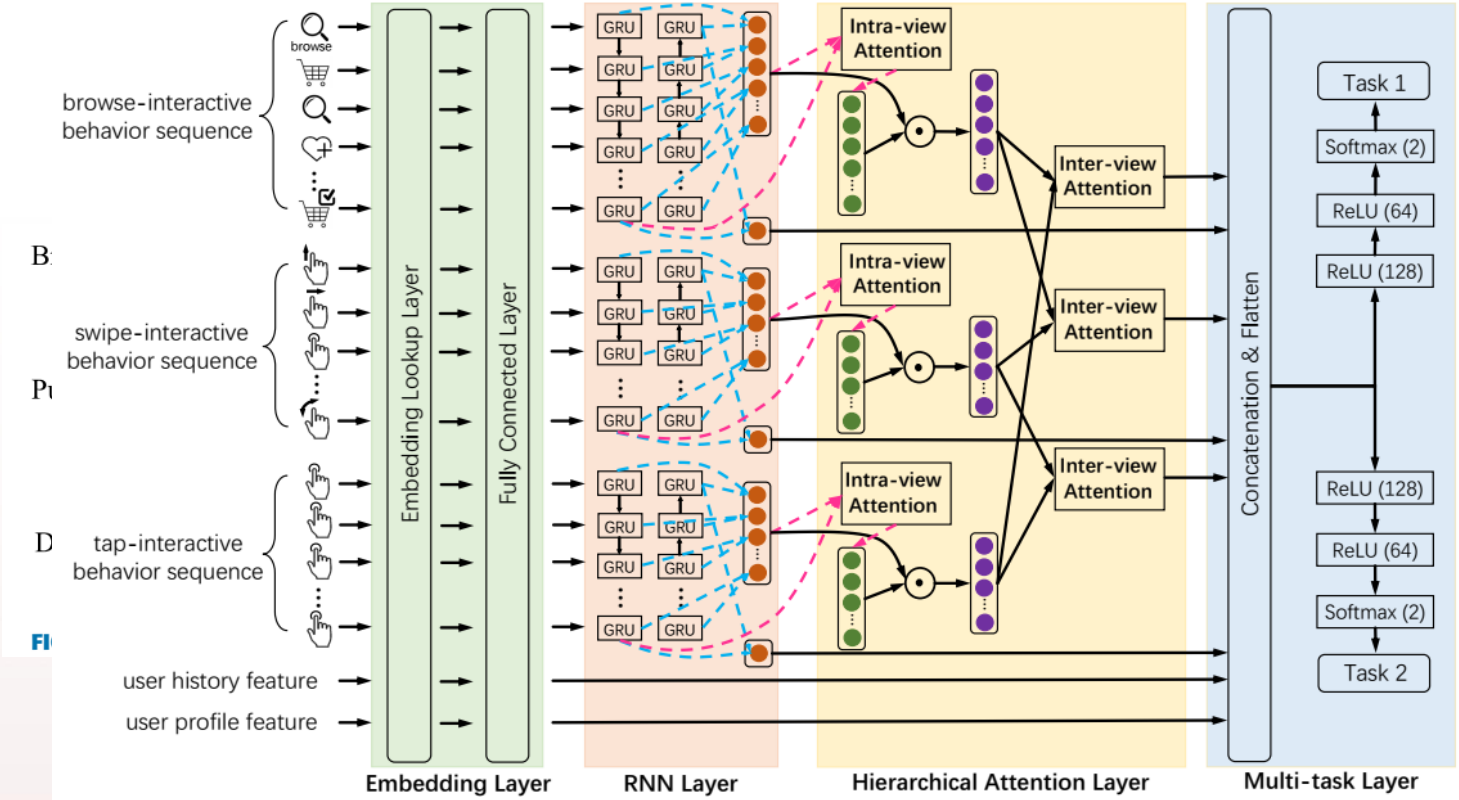


Figure 1: The model architecture of DIPN.

[1] Ling C, Zhang T, Chen Y. Customer purchase intent prediction under online multi-channel promotion: A feature-combined deep learning framework[J]. IEEE Access, 2019, 7: 112963-112976.

[2] Guo L, Hua L, Jia R, et al. Buying or browsing?: Predicting real-time purchasing intent using attention-based deep network with multiple behavior[C]//Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 2019: 1984-1992.

03 在线旅游购买预测模型

三、在线旅游购买预测模型

1. 数据驱动：

- 数据来源

  本文所使用的旅游产品数据集来自于中国在线旅游平台(Online Travel Agency, OTA):途牛旅游网[1] (合作方式获取).

- 途牛提供的原始数据主要分为3大类:

用户	会话		日期	# 点击流	# 会话	# 用户	# 购买会话	城市	搜索内容
U1	*S1*	201	2013.08	77,965,634	8,576,557	13,940,363	181,083	安	Null
U1	*S1*	201	2013.07	83,906,441	11,558,465	14,155,555	187,918	安	西安一日游
...	...		2012.12	26,352,279	6,424,642	4,844,989	56,140	...	...
U2	*S2*	201	2012.11	20,224,856	5,740,788	4,457,294	15,635	京	Null
U2	*S2*	201	2012.10	20,558,671	5,564,276	4,135,774	17,550	京	厦门游
U2	*S3*	201	2012.09	33,909,679	7,862,433	5,572,967	32,065	京	Null
U2	*S3*	201	2012.08	34,250,002	7,852,153	5,596,534	29,904	京	鼓浪屿
U2	*S3*	201	2012.07	39,351,058	9,432,774	6,654,679	29,096	京	Null

注释：购买会话是通过用户所属会话包含的“预订”页面推断的。

三、在线旅游购买预测模型

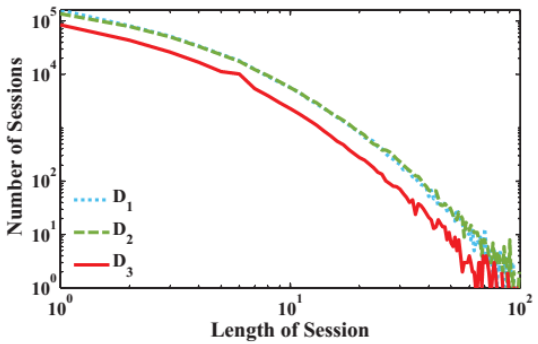
1. 数据描述：

- ✓ D1 (Summer Vacation: 1 to 7 Aug., 2012)
- ✓ D2 (National Day: 24 to 30 Sep., 2012)
- ✓ D3 (Slow Season: 1 to 7 Nov., 2012)

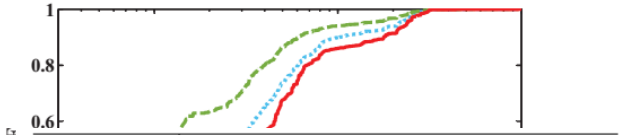
数据集	时间	# 记录	# 用户	# 会话	# 订单	订单转化率
D_1	2012 08.01-08.07	2,022,633	364,067	431,321	7,284	1.69%
D_2	2012 09.24-09.30	1,980,299	341,878	403,032	10,236	2.54%
D_3	2012 11.01-11.07	941,930	190,292	217,692	2,731	1.26%

2.旅游数据分析：

- 长尾现象显著:
- 类不均衡以及冷启动问题显著:

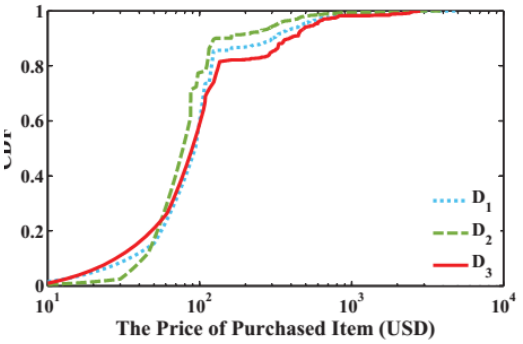


(a) 会话长度的分布



数据集	# 冷启动用户	比例 (%)
D_1	225,749	52.34
D_2	245,703	60.96
D_3	121,006	55.61

(b) 所有产品价格



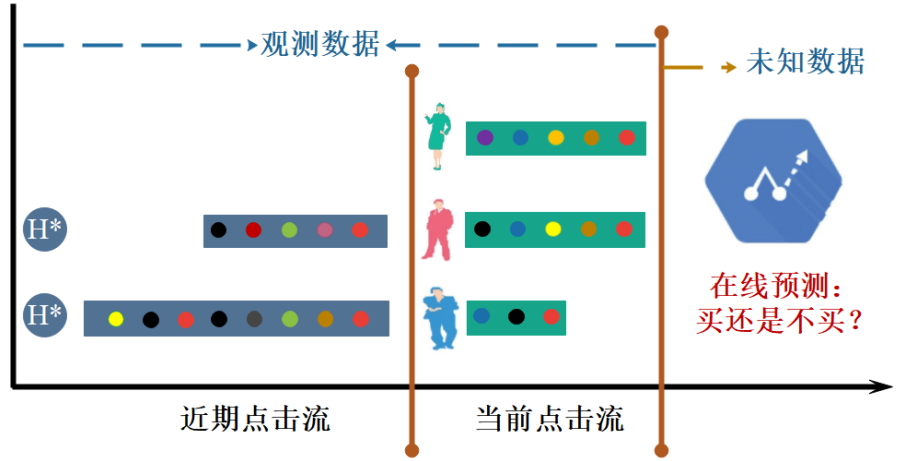
(c) 购买产品价格

三、在线旅游购买预测模型

途牛旅游网 Tuniu.com

3.在线旅游用户细分：

管理学领域很多研究已经证实了一刀切(one-size-fits-all) 的营销策略远远不是最佳的实践方案.



- ✓ **First-Time Visitors**:此类用户首次访问该网站,这意味着没有一个历史信息可以去了解这些用户,除了当前点击流.
- ✓ **Ever-Visited Users**:此类用户曾经访问过网站，但是在近期没有访问记录,例如,在最近的一个月点击流,所以除了一些人口统计信息,仅有当前的点击流可用.
- ✓ **Recent-Visited Users**:此用户非常活跃,即:他们在近期访问了网站.因此,近期的和当前的点击流,还有一些人口统计信息是可用的.

三、在线旅游购买预测模型

4. 变量定义（用于购买预测模型）

- ✓ 当前点击流变量:

  针对First-Time Visitors，

  Ever-Visited Users，

  Recent-Visited Users.

- ✓ 近期点击流变量:

  只针对Ever-Visited Users，

  Recent-Visited Users.

- □ 特点:

  融入了当前会话特征，用户近期（历史）行为特征，用户信息（是否会员），旅游领域信息（如旅游时空因素）等.

	特征（变量）		描述	p-value	
近期点击流特征		用户统计信息			
	1	$Member_i$	用户 i 是否为会员 (1=yes, 0=no)	***	✓
		点击流度量			
	2	$LVDays_i$	距离用户 i 最近一次访问间隔天数 (Sigmoid)	***	✓
	3	$TotV_i$	用户 i 在近一个月总的访问次数	***	✓
	4	$PAvgV_i$	用户 i 近一个月在平台浏览产品的平均价格（单位：美元）	**	✓
	5	$PDevV_i$	用户 i 近一个月在平台浏览产品价格（单位：美元）的标准差	**	
	6	$LDwell_i$	用户 i 近一个月内最近的一次会话的停留时间（单位：秒）	*	✓
	7	$DwellAvg_i$	用户 i 近一个月内访问的所有会话的平均停留时间（单位：秒）	*	✓
		购买行为			
	8	$TotP_i$	用户 i 近一个月内累计购买的总次数	***	✓
	9	$LPDays_i$	用户 i 距离近一个月内最近的一次购买的时间间隔（单位：天）(Sigmoid)	***	✓
	10	$MAvgP_i$	用户 i 近一个月内购买旅游包的平均花费（单位：美元）	***	✓
当前点击流特征		点击流度量			
	11	$PAvg_j$	当前会话 j 中浏览旅游包的均价（单位：美元）(Log)	***	✓
	12	$PDev_j$	当前会话 j 中浏览旅游包价格（单位：美元）的标准差 (Log)	**	✓
	13	$Length_j$	会话 j 的长度 (Log)	***	✓
	14	$Dwell_j$	当前会话 j 的停留时间（单位：秒）(Log)	***	✓
	15	$Search_j$	当前会话 j 中使用的搜索引擎类型 (0=none, 1=offsite, 2=onsite)	***	✓
	16	$TRegions_j$	当前会话 j 中旅游目的地分布的熵	***	
	17	$PTypes_j$	当前会话 j 中旅游类型分布的熵	**	
	18	$RPages_j$	当前会话 j 中包含旅游包相关展示页面的比例	***	
		旅游时空度量			
	19	$Location_j$	用户 i 所在城市和会话 j 旅游包出发城市的距离相似度	***	
	20	$Holiday_j$	用户 i 当前会话 j 的时间与最近节假日时间戳的间隔天数	**	
	21	$Weekend_j$	当前会话 j 的所在日期是否为周末	**	✓

注释：(1) “Sigmoid” 表示将变量（特征）通过 Sigmoid 函数 $\frac{1}{1+e^{-x}}$ 规格化到区间 $[0,1]$；
(2) “Log” 表示将变量（特征）通过对数函数 $\log_{10} x$ 计算取自身对数；
(3) “曼 -惠特尼秩和检验” (Mann-Whitney U test)[158]：*<0.05; **<0.01; ***<0.001。

三、在线旅游购买预测模型

5. 在线购买模式分析

✓ 用户类型对订单转化率(CR) 的影响

数据集	First-Time Visitors	Ever-Visited Users	Recent-Visited Users	
			未买	买
D_1	0.88%	2.55%	2.19%	17.15%
D_2	1.34%	1.39%	4.47%	17.70%
D_3	0.61%	1.73%	1.92%	17.60%

✓ 搜索行为(站内搜索和站外搜索)对CR的影响

站内搜索：Tuniu内部搜索引擎

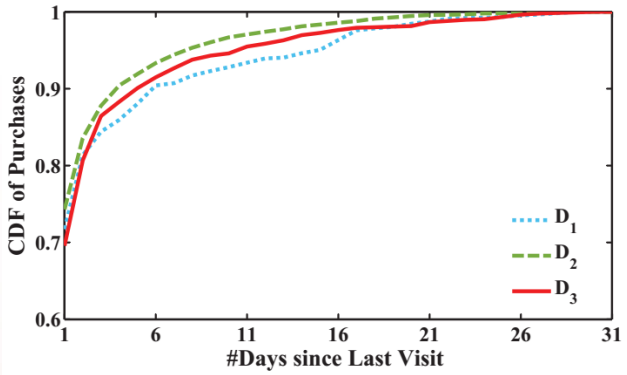
站外搜索：通过外部搜索引擎跳转到平台页面的（如Baidu，360和Google等）

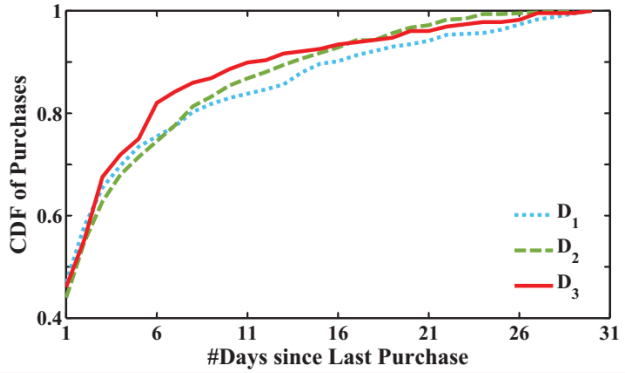
		D_1		D_2		D_3	
		站内	站外	站内	站外	站内	站外
$P(Y\|X=1)$,	买	95.12%	3.74%	96.23%	4.19%	94.78%	4.69%
$Y=\{0,1\}$	未买	4.88%	96.26%	3.77%	95.81%	5.22%	95.31%
$P(X=1\|Y)$,	买	21.76%	61.47%	25.61%	53.25%	20.39%	59.65%
$Y=\{0,1\}$	未买	1.67%	56.83%	1.78%	49.37%	1.22%	58.89%

三、在线旅游购买预测模型

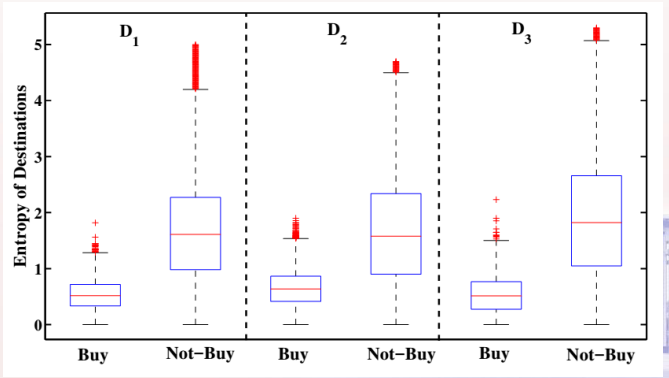
途牛旅游网 Tuniu.com

✓ 近期访问和购买对于CR的影响



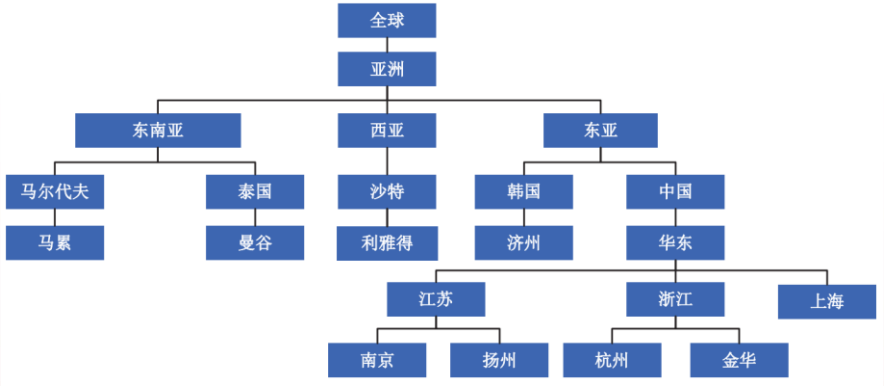


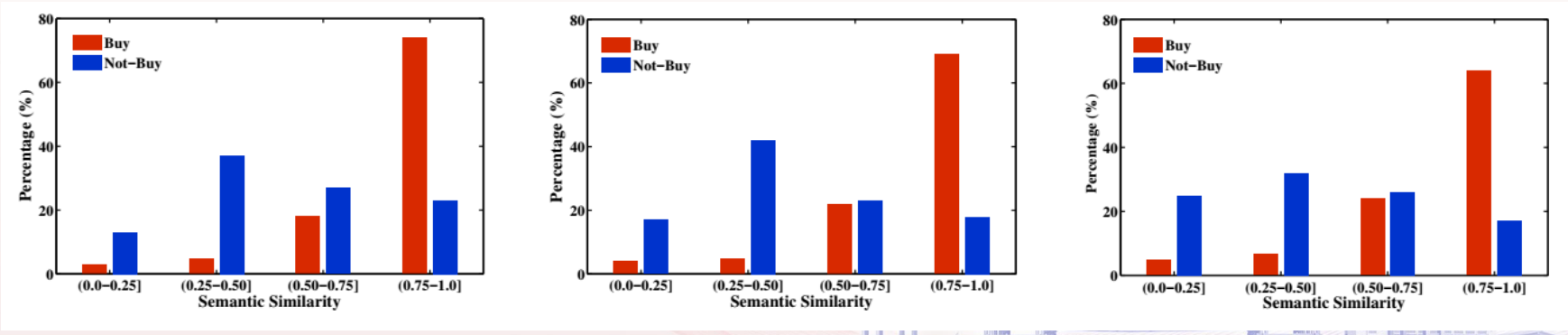
✓ 旅游包区域分布的熵对于CR的影响



三、在线旅游购买预测模型

- ✓ 地理信息相似度对于CR的影响为此,我们使用树的结构相似性来定义两个城市之间的语义关系.具体而言,我们充分利用来自于United Nations geoscheme[2]的行政划分去构建一个该旅游数据中包含的所有的线路出发城市和用户居住城市的地理树.然后,城市之间的距离是转换为两个节点在这个地理树上的相似性.





[2] https://en.wikipedia.org/wiki/United_Nations_geoscheme

三、在线旅游购买预测模型

6. co-EM-LR (co-EM Logistic Regression)预测模型

$$h_\theta(x_i) = \frac{1}{1+\exp(-\theta^\top x_i)} = \frac{1}{1+\exp(-\theta_1^\top x_{i1} - \theta_2^\top x_{i2})}. \tag{3.4}$$

$$\mu_y = \frac{1}{|\mathcal{D}_l^y| + \lambda|\mathcal{D}_u^y|}\Big(\sum_{x_i \in \mathcal{D}_l^y} h_\theta(x_i) + \lambda \sum_{x_j^* \in \mathcal{D}_u^y} h_\theta(x_j^*)\Big), \tag{3.5}$$

$$\sigma_y^2 = \frac{1}{|\mathcal{D}_l^y| + \lambda|\mathcal{D}_u^y|}\Big(\sum_{x_i \in \mathcal{D}_l^y} (h_\theta(x_i) - \mu_y)^2 + \lambda \sum_{x^* \in \mathcal{D}_u^y} (h_\theta(x_j^*) - \mu_y)^2\Big), \tag{3.6}$$

$$\begin{aligned} p(\hat{y}_j|x_j^*; \theta, \mu_{\hat{y}_j}, \sigma_{\hat{y}_j}^2) &= \frac{p(x_j^*|\hat{y}_j)p(\hat{y}_j)}{\sum_{y\in\{1,0\}} p(x_j^*|y)p(y)} \\ &= \frac{\mathcal{N}(h_\theta(x_j^*)|\mu_{\hat{y}_j}, \sigma_{\hat{y}_j}^2)p(\hat{y}_j)}{\sum_{y\in\{1,0\}} \mathcal{N}(h_\theta(x_j^*)|\mu_y, \sigma_y^2)p(y)}, \end{aligned} \tag{3.8}$$

$$l(\theta) = \log\Big(\prod_{x_i \in \mathcal{D}_l} p'(y_i|x_i;\theta)^{\Lambda_i} \prod_{x_j^* \in \mathcal{D}_u'} p'(y_j^*|x_j^*;\theta)^{\Lambda_j}\Big), \tag{3.10}$$

$$p'(y_i|x_i;\theta) = (h_\theta(x_i))^{y_i}(1-h_\theta(x_i))^{1-y_i}. \tag{3.9}$$

Algorithm 1 co-EM-LR 算法

Input: $\mathcal{D}_l$：标记数据集；$\mathcal{D}_u$：无标记数据集；

Output: 每一个样本 $x_j^* \in \mathcal{D}_u$ 的概率值 $p'(y_j|x_j^*)$, $y_j \in \{1,0\}$；

在 $\mathcal{D}_l$ 上通过视图 V_2 训练一个初始的 LR 模型，同时生成 θ_2；
repeat
  for $v = 1$ to 2 **do**
    通过计算在互补视图 $V_{\bar{v}}$ 上的 $h_{\theta_{\bar{v}}}(x_j^*)$ 获取 $\mathcal{D}_u^1$；$\mathcal{D}_u^0 = \mathcal{D}_u \setminus \mathcal{D}_u^1$；
    通过式 (3.5) 和式 (3.6) 估算参数 μ_1, μ_0, σ_1^2 和 σ_0^2；
    依据式 (3.8) 并通过参数 $\theta_{\bar{v}}$ 估算 $p(\hat{y}_j|x_j^*)$，$\forall x_j^* \in \mathcal{D}_u$；
    生成一个伪标记数据集 $\mathcal{D}_u'$；
    使用一个带有平滑因子 Λ_j 的 SGD 方法，并通过最大化式 (3.10) 更新参数 θ_v；
  end for
until 条件收敛
return 通过式 (3.9) 计算概率值 $p'(y_j|x_j^*)$, $y_j \in \{1,0\}$, $\forall x_j^* \in \mathcal{D}_u$；

- 模型特点:
  - ✓ 半监督学习:充分利用大量无标记样本,提升学习性能.
  - ✓ 多视图学习:不同视图的变量协同工作得出一致性的决定.
  - ✓ 基础学习方法为逻辑回归:模型的可解释性较强.

三、在线旅游购买预测模型

7000个恶意用户中，
有6954个被正确分类为“恶意用户”，
有46个被错误分类为“非恶意用户”

7.评价指标

混淆矩阵

		Predicted Class		合计
Actual Class		Class=购买用户	Class=非购买用户	
	Class=购买用户	TP=2588	FN=412	3000
	Class=非购买用户	FP=46	TN=6954	7000
	合计	2634	7366	10000

Yes 类的评价

$$\text{Precision (P)} = \frac{TP}{TP+FP} = 98.3\%$$

$$\text{Recall (R)} = \frac{TP}{TP+FN} = 86.3\%$$

$$\text{F-measure (F)} = \frac{2PR}{P+R} = 91.9\%$$

具有高确率的分类器，非对角线的项应接近于零

The harmonic mean of *R* and *P*!

三、在线旅游购买预测模型

途牛旅游网 Tuniu.com

8. 整体性能对比

- ✓ 在性能指标***R***上,co-EM-LR是优于其他基准方法,但是在性能指标***P***上,略微优于其他基准算法.然而,从性能指标***F***来看co-EM-LR的整体性能优于其他基准算法的.co-EM-LR模型比起排名第二的基准算法在性能指标***R***上提升了**14.6%-28.3%**,在性能指标***F***上提升了**5.6%-11.5%**.
- ✓ **Recent-Visited Users**上的性能是最好的，其次是**Ever-Visited Users** 和 **First-Time Visitors.**

评价指标	方法	D_1				D_2				D_3			
		U_1	U_2	U_3	Overall	U_1	U_2	U_3	Overall	U_1	U_2	U_3	Overall
P	co-EM-LR	0.857	0.923	0.947	0.927±0.006	0.835	0.861	0.874	0.869±0.006	0.845	0.896	0.913	0.899±0.005
	co-EM	**0.951**	**0.972**	**0.986**	**0.974**±0.007	0.876	0.931	0.956	0.920±0.008	0.861	0.894	0.925	0.887±0.009
	co-Training	0.827	0.896	0.926	0.864±0.013	0.766	0.852	0.883	0.843±0.012	0.879	0.943	0.978	0.952±0.011
	LR	0.751	0.894	0.912	0.879±0.028	0.853	0.882	0.913	0.861±0.020	0.793	0.836	0.886	0.882±0.021
	RF	0.891	0.944	0.977	0.962±0.019	**0.948**	0.979	0.985	0.980±0.016	**0.965**	**0.978**	**0.982**	**0.975**±0.021
	GBDT	0.912	0.971	0.978	0.969±0.011	0.941	**0.980**	**0.987**	**0.982**±0.012	0.948	0.974	0.979	0.973±0.014
	Average	0.865	0.933	0.954	-	0.870	0.914	0.933	-	0.882	0.920	0.944	-
R	co-EM-LR	**0.357**	**0.462**	**0.568**	**0.493**±0.005	0.423	**0.547**	**0.719**	**0.643**±0.008	0.358	**0.569**	**0.716**	**0.605**±0.009
	co-EM	0.181	0.201	0.221	0.196±0.009	0.198	0.291	0.301	0.281±0.008	0.182	0.217	0.230	0.203±0.009
	co-Training	0.303	0.383	0.431	0.356±0.009	**0.542**	0.476	0.578	0.492±0.011	0.298	0.337	0.349	0.321±0.009
	LR	0.271	0.351	0.421	0.336±0.022	0.372	0.461	0.487	0.424±0.019	0.291	0.378	0.418	0.347±0.021
	RF	0.328	0.411	0.441	0.394±0.015	0.365	0.468	0.521	0.465±0.019	0.381	0.473	0.489	0.501±0.018
	GBDT	0.335	0.429	0.481	0.422±0.016	0.391	0.502	0.545	0.501±0.014	**0.427**	0.521	0.553	0.528±0.017
	Average	0.296	0.373	0.427	-	0.382	0.458	0.525	-	0.323	0.416	0.460	-
F	co-EM-LR	**0.509**	**0.618**	**0.717**	**0.644**±0.005	0.563	0.670	**0.789**	**0.739**±0.006	0.507	**0.697**	**0.809**	**0.723**±0.007
	co-EM	0.304	0.333	0.361	0.326±0.007	0.323	0.443	0.458	0.431±0.007	0.300	0.349	0.367	0.330±0.009
	co-Training	0.444	0.537	0.588	0.504±0.012	**0.587**	**0.689**	0.709	0.660±0.011	0.445	0.497	0.514	0.480±0.010
	LR	0.398	0.504	0.576	0.486±0.017	0.518	0.606	0.635	0.568±0.014	0.426	0.521	0.568	0.498±0.012
	RF	0.479	0.573	0.608	0.559±0.018	0.438	0.520	0.581	0.631±0.019	0.429	0.503	0.565	0.635±0.018
	GBDT	0.490	0.595	0.645	0.588±0.012	0.552	0.664	0.702	0.663±0.015	**0.583**	0.679	0.705	0.685±0.015
	Average	0.437	0.527	0.583	-	0.497	0.599	0.646	-	0.448	0.541	0.588	-

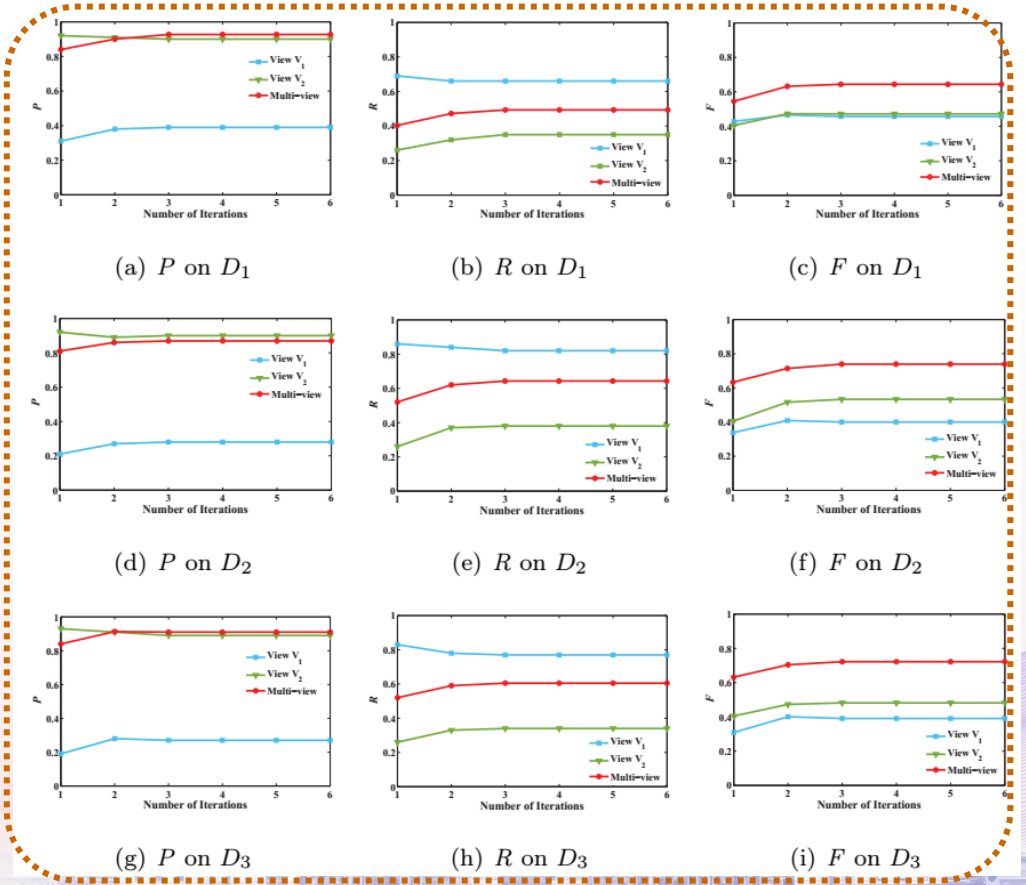
注释：U_1、U_2 和 U_3 分别代表 first-time visitors、ever-visited users 和 recent-visited users。

注:在购买预测的场景中***R***比***P***更加重要,想象每天成千上万的用户访问了电子商务网站,呼叫中心的团队希望呼叫所有人以提升访问到购买的转化率,这显然是不可能的任务.因此,希望最有可能成为购买者的在线用户是他们选择呼叫的目标客户.从这个角度来看,co-EM-LR具有更高的***R***值对于电子商务平台更有价值和意义。

三、在线旅游购买预测模型

9. 多视图学习的影响

- ✓ 两个视图是相互迥异和补充的:
  视图 V_1 通常带来低的***P***值,但是较高的***R***值,这意味着分类器在V_1视图上倾向于标记大量的"即将购买"的会话.而模型在当前点击流的V_2视图上更加保守,它更加倾向于产生一列更有可能买的会话.然而,多视图学习的***P***和***R***曲线是基于V_1和V_2中间的.这意味着co-EM-LR充分利用了多视图学习去平衡了两个单独视图上的准确度***P***和召回度***R***,最后提升了整体分类器的精准度.
- ✓ co-EM-LR能够快速收敛:
  所有的曲线在第二轮迭代有一个快速的增长(下降),然后在收敛之前保持基本持平.第一轮迭代之后的增长是由于互补效应,模型在第二轮首次感知到了该效应.在实践中,我们设置了theta变化的阈值为0.0001,co-EM-LR模型通常在迭代10轮左右即可收敛.



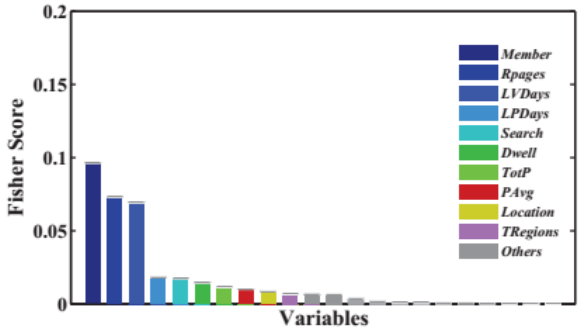
(a) P on D_1 (b) R on D_1 (c) F on D_1

(d) P on D_2 (e) R on D_2 (f) F on D_2

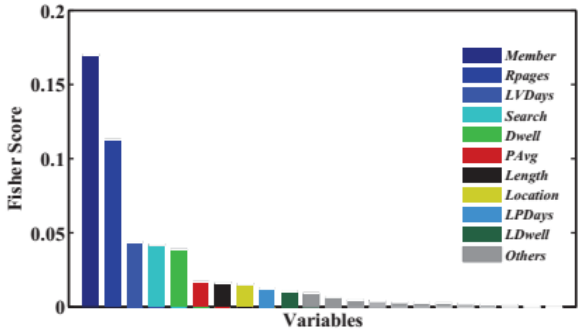
(g) P on D_3 (h) R on D_3 (i) F on D_3

三、在线旅游购买预测模型

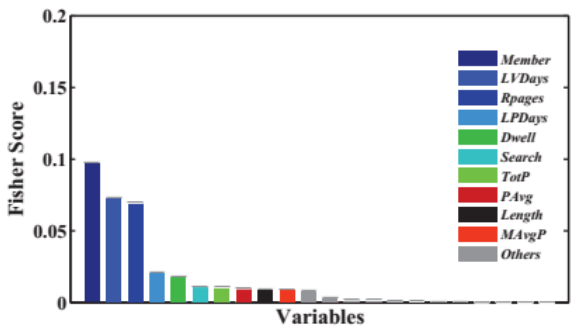
10. 变量重要性度量



(a) D_1



(b) D_2



(c) D_3

彩色竖条代表最重要的10个变量,灰色竖条代表拥有更低Fisher score值的变量.

✓在3个数据集上最重要的10个变量是非常相似的.

✓此外,3个数据集上最重要的3个特征不是*Member*→*LVDays*→*Rpages* 就是*Member*→*Rpages*→*LVDays*,这意味着用户个人信息,当前会话包含旅游产品页面的比例,近期会话的访问时间是显著影响在线购买决策的关键性因素.

✓当前会话的浏览行为(如*Search*和*Dwell*),近期会话浏览的行为(如*LPDays*),金钱的花费(如*Pavg*)和用户的地理信息(如*Location*)均对在线购买决策有着显著的影响.

谢谢观赏 下节课见