高潜用户预测实践

项目：机器学习期末考核
姓名：杜圣辉
学号：2017111136
班级：17 级信息管理与信息系统 1 班

工具：python(3.6), Jupyterlab, MySQL(8.0.20), Navicat

一、 任务背景与数据集介绍

第一节 任务背景

1 背景

电子商务的快速发展带来了海量的数据，在战略和运营层面为公司带来诸多挑战。高潜用户，即在未来一段时间有高潜在购买行为的用户。如何从海量数据中提取有价值的信息并预测海量消费者中未来有高购买意向者，并基于此进行更进一步的营销、运营手段，是整个电子商务领域解决精准营销问题的关键。

正好在 2020 年 INFORMS 举办的 MSOM Research Challenge 中，京东提供了他们脱敏的用户行为数据，能够作为我实验的基础。此数据集提供了 2018 年 3 月份某一特定产品类别超过 3 万 sku 的 250 万客户的信息，并且信息、维度等较为全面，且所有的行为都发生在某一特定品类（category）之下，能够更好让我的实验聚焦和提取特征。

同时，在学习机器学习的过程中，我借势学习了推荐系统的相关知识，尝试从 CTR 领域常用模型出发,进行上一层的高潜用户预测实践,将近期所学的模型中最有代表性和意义，同时较为友好者实现，并比较其在高潜预测场景下效果。

2 高潜预测问题定义与数据集构建

如前面所说，高潜用户，即在未来一段时间有高潜在购买行为的用户。

通过调研现在业内的方法和一些过往的市场营销学资料,发现目前广泛采用的是根据用户在一定时间内的行为，来预测其在未来几天内是否会购买。以过去某段时间内的数据，选择用户、提取特征和标签，送入模型进行训练后，并在业务内的“当天”进行离线训练，测试，并检验模型的有效性，其输出可以根据用户的转化概率分配不同的运营策略（如针对摇摆用户发券，或者进行 A/B 实验，基于因果推断，利用 Uplift 或概率图的方法找到对消费券

最敏感的用户，进行消费券发放[1])，也可以作为输入训练相关特征。但不论如何，高潜预测模型在业务场景中都是十分重要的节点。

基于此，为了更贴合实际场景，我选择利用不同时间窗口内采集的数据分别作为训练集、测试集：以 3 月 10 日至 27 日的时间窗口为数据集，其中 10~23 日（14 天）统计用户特征，24~27 日（4 天）提取用户标签，即有购买行为则标记为 1，否则为 0。同理编辑 3 月 13 日至 30 日的时间窗口内数据为测试集。（时间窗口选择详情见三，第一节，第一部分）

第二节 数据集介绍

1 数据来源

数据集来自 2020 年 MSOM Data Driven Research Challenge 中提供的数据集，由于数据量较大，我上传至了百度云：

链接：https://pan.baidu.com/s/1Ffmh4a8-0mVv3EtakrEvTA

提取码：2uw3

2 表介绍与基本探索

京东提供的数据包含 2.5 万消费者信息和 3 万余个 sku 信息，以及他们中的一部分在 2018 年的行为记录（千万条点击、购买记录等），同时举办方也提供了仓储信息，以支持多维度的分析。

本次我将用到的表有用户表（user）、点击日志（click）、订单表（order）。数据存储至本地 MySQL 数据库中，以实现更快和方便的表处理和清洗。同时，为了避免表和 MySQL 中关键词冲突，我会在表明前加入“jd”，如 jduser，用作区分。

中间表和临时表不在本次介绍当中，所有的表生成方法我会作为附件放在文件中。

接下来对数据表的字段、取值类型和分布做基本介绍：

2.1 用户表（jduser）

用户表描述了 2018 年 3 月期间购买了至少一个特定类别 sku 的 457,298 名用户的每个特征。对于每个重复客户，将根据其过去的购买行为对相应的用户进行分类，user_level 的值为 0、1、2、3 或 4，其中较高的 user_level 与较高的过去总购买价值相关联。对于企业用户(例如农村地区的小商店或小型企业)，相应的 user_level 的值为 10。然而，对于首次购买者，user_level 为-1。用户所在城市的等级 city_level 根据用户的常用收货地址得出，该字段的值为 1-5。在这里，一级对应的是北京、上海等一线城市；二级城市为省会城市；3 - 5 级城市是较小的城市;如果没有数据，则值为−1。

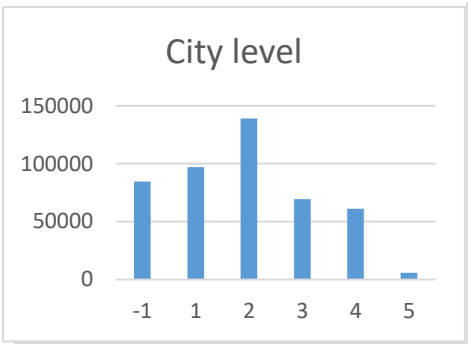
京东的客户在购买时不需要提供任何人口统计信息，但是京东拥有一个复杂的人工智能系统来估计用户的人口统计数据（demographic features）。所以在用户表中还包含以下的估计信息：

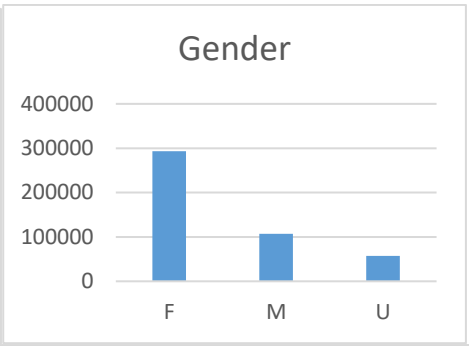
(a)性别：F:女性，M:男性，U:未知；

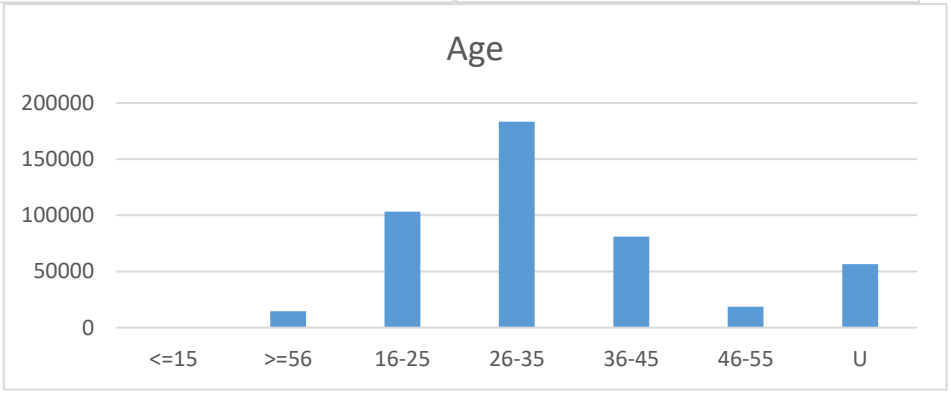
(b)年龄：<=15，16-25，26-35，36-45，46-55，>=56，U:未知；

[1] 阿里文娱技术，阿里文娱智能营销增益模型（Uplift Model）技术实践，2020-03-25

(c)婚姻: 使用者的婚姻状况(M:已婚，S:单身，U:未知);
(d)教育程度: 1:高中以下，2:高中毕业或同等学历，3:学士学位，4:研究生学位，−1:未知;
(e)购买力： 1 - 5, 1 为最高购买力; -1，没有估计;





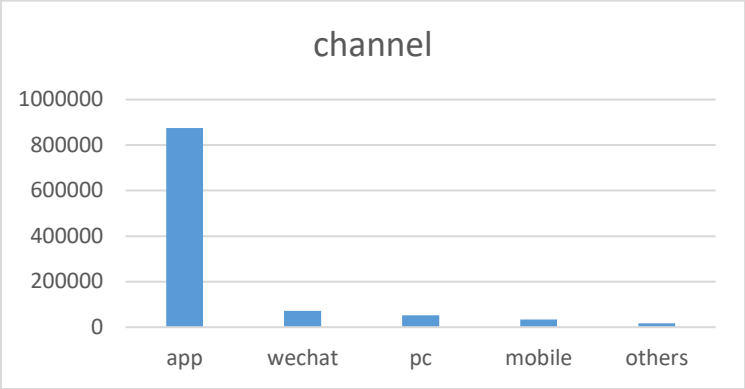


字段名	数据类型	说明	示例
user_ID	string	用户编号	000000f736
user_level	int	用户等级	10
first_order_month	string	用户在京东首次购买行为的月份	2017-07
plus	int	是否是 plus 会员	0
gender	string	性别（估计）F 女,M 男,U 未知	F
age	string	年龄（估计）	26–35
marital_status	string	婚姻状况（估计）	M
education	int	User edu 教育水平（估计）	3
purchase_power	int	用户购买力（估计）	2
city_level	int	城市级别	1

2.2 点击日志（jdclick）

点击表通过用户浏览历史建立用户和 sku 之间的链接。表中的每个条目表示用户在特定 SKU 页面上的“点击事件”。但这个表包含的用户不仅来自用户表中确定的购买了至少一个 SKU 的用户，而且来自没有完成购买订单的“其他用户”，这也对后续的 user_sku 匹配并预测购买行为带来困难。

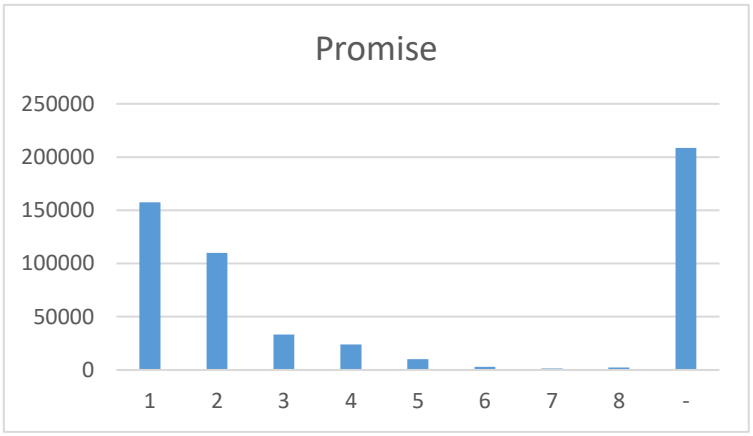
在 channel 字段这里，数据被分为五个字符串值:pc、mobile、app、微信和 other。图 10 总结了所有通道中所有单击事件的分布情况。由于智能手机在中国的普及和移动支付选项(如微信支付)的普及，大部分的点击事件来自 app 和微信频道。

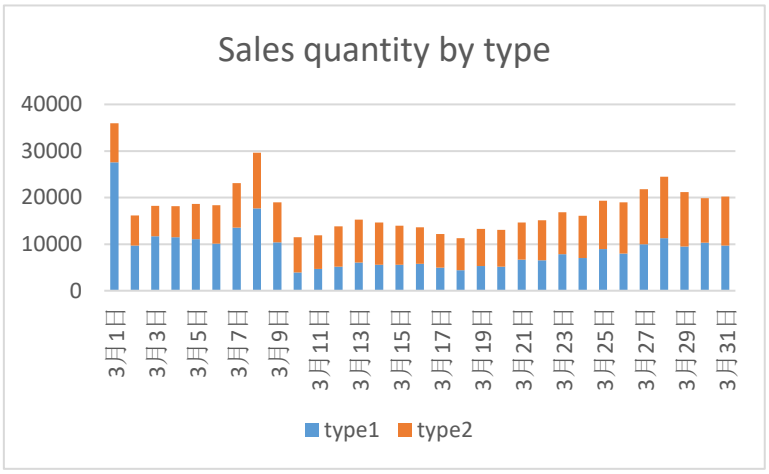


字段名	数据类型	说明	示例
sku_ID	string	商品编号	b4822497a5
user_ID	string	用户编号	94ff800585
request_time	string	点击行为发生时间	2018-03-01 23:57:53
channel	string	点击渠道	wechat

2.3 购买日志（jdorder）

订单表包含了 2018 年 3 月此类别中的 486,928 个用户订单。从 promise 图可以看出，大多数订单都承诺在 2 天内交货。按订单类型划分订单数量，可以发现 type1 type2 较为平衡。





订单表提供了每个 SKU 的产品定价和促销活动信息。对于每个条目，我们在字段 original_unit_price 中表示 SKU 的原始列表价格，客户为 SKU 实际支付的价格为 final_unit_price。一个 SKU 在任何给定时刻对所有客户的原始标价都是相同的，但由于各种折扣或促销，最终的价格会因客户的不同而不同。原始价格和最终价格之间的“差距”代表了与每个 SKU 不同的促销活动和折扣。京东平台上常见的促销折扣有四种类型:(1)SKU 直接折扣，(2)集团促销，(3)捆绑促销，(4)礼品。这四种类型的折扣可以描述如下:(1) SKU 的卖方可以提供直接折扣的降价。这个折扣反映了在产品细节页面上所述的清单价格的降低。(2)集团促销：SKU 的卖家可以提供数量折扣，吸引客户购买更多。本次数量折扣促销可以采取“满 199 减 100”、“买 3 送 1”等多种形式。 (3)如果客户在订单中购买了“预先指定的套装”sku，卖方可以提供捆绑促销。(4)如果顾客购买了“预先指定的一组”SKU，卖方可能会提供一个 SKU 作为“免费礼物”(gift_item value = 1)(例如，购买 10 支铅笔和 20 张纸，可以得到一个免费橡皮擦)。每个礼品的最后单价总是等于 0。

字段名	数据类型	说明	示例
order_ID	string	订单编号	3b76bfcd3b
user_ID	string	用户编号	3cde601074
sku_ID	string	SKU 编号	443fd601f0

order_date	string	下单日期(yyyy-mm-dd)	2018-03-01
order_time	string	下单时间 (yyyy-mm-dd HH:MM:SS)	2018-03-01 11:10:40.0
quantity	int	购买数量	1
type	int	自营或第三方	1
promise	int	保证送达日期	2
original_unit_price	float	原价	99.9
final_unit_price	float	最终购买价	53.9
direct_discount_per_unit	float	直接折扣	5.0
quantity_discount_per_unit	float	集团促销	41.0
bundle_discount_per_unit	float	捆绑促销	0.0
coupon_discount_per_unit	float	优惠券	0.0
gift_item	int	礼物	0

4 衍生表

3.1 标签表（jd_high_user_base_train/test）

本表对训练集、测试集中选取用户进行打标签操作，如果在打标签的窗口内有过购物行为，则标记为 1，否则为 0。

字段名	数据类型	说明	示例
User_ID	string	用户编号	007831ead9
Label	int	标签	1

3.2 用户画像表(jd_hish_user_train/test_bhav_protrait)

记录训练、测试集中用户的行为维度画像表[2]。它能够刻画用户过往一段时间内在此品类下的访问习惯。这里会记录用户 1、3、7、14 天内的点击、消费次数。同时会记录点击、消费的时间衰减因素（详见下一部分第一节的时间衰减部分）。

字段名	数据类型	说明	示例
user_ID	string	用户编号	e8b625a7cf
CLK_1D	Int	过去 1 天内总点击次数	232
CLK_3D	Int	过去 3 天内总点击次数	247
CLK_7D	Int	过去 7 天内总点击次数	258
CLK_14D	Int	过去 14 天内总点击次数	398
CLK_DISTRACT_SCORE	float	订单时间衰减分	153.0453
ODR_1D	Int	过去 1 天内总下单次数	2
ODR_3D	Int	过去 3 天内总下单次数	2
ODR_7D	Int	过去 7 天内总下单次数	2
ODR_14D	Int	过去 14 天内总下单次数	2
ODR_DISTRACT_SCORE	float	订单时间衰减分	1.1699

[2] 赵宏田，用户画像——方法论与工程化解决方案，机械工业出版

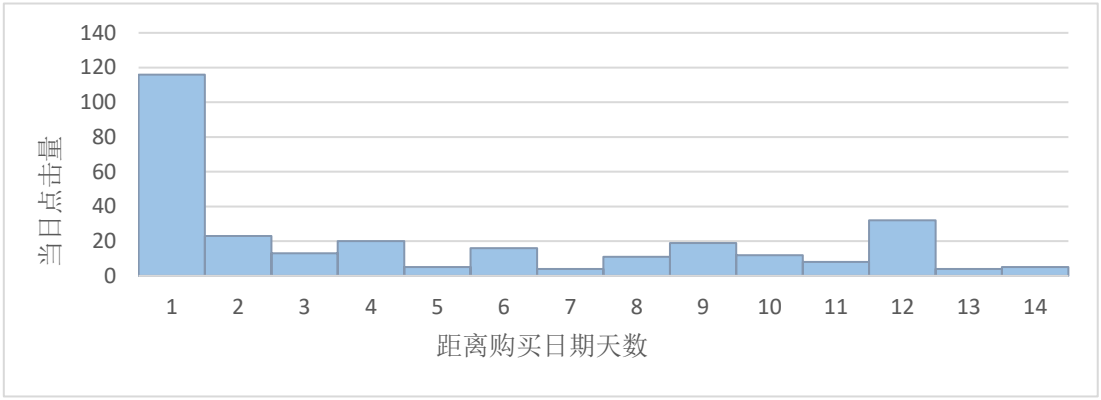
二、所采取的方法

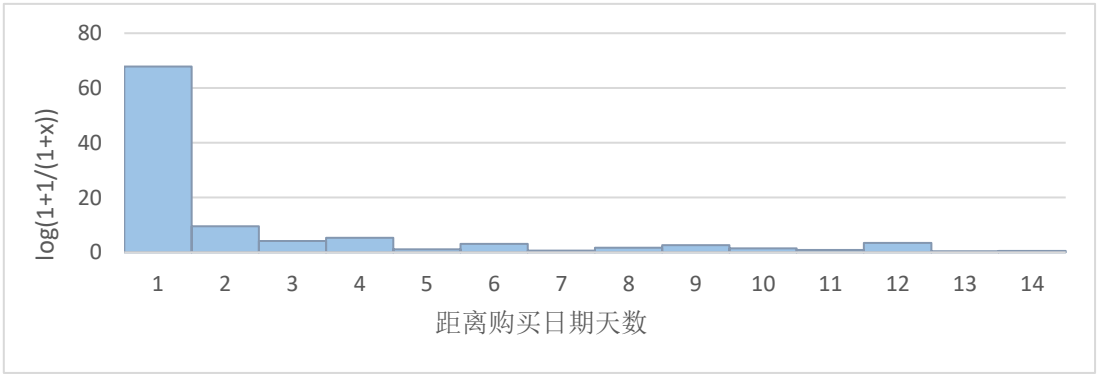
第一节 特征处理

1 时间衰减

根据常识经验我们可以知道，过往的行为会影响我们未来的决策，但是这一影响往往会随着时间的流失而逐渐衰减。这里的时间衰减因素就是为了反应这种现象。

这里我尝试用 $w = \log_2(1 + \frac{1}{1+x})$ 进行刻画这一影响权重，其中 x 表示距今日期。那么可知 x 一定是非负数，那么 log 函数的值域就落在区间 [1,0) 之间，且随着距今日期 x 的增加而递减，同时，对于每天的行为数据乘以权重，就能够得到当天的行为分，如下图所示：





如果对用户过往时间内的衰减行为进行累加，就得到了点击、订单的时间衰减行为分：

$$\text{score}_i = \sum_{i=1}^{14} y_{x,i} * \log_2(1 + \frac{1}{1+x})$$

其中，i 表示为对应的$user_i$，$y_{x,i}$表示用户 i 在距今 x 天当天的行为次数，并且这里的对数函数取以 2 为底进行计算。

第二节 模型

在本次实践中，我一共使用了三种有效模型。其中：

（1）.传统机器学习模型包括作为 Baseline 模型的逻辑回归（Logistic Regression），以及在此之上，由 Facebook 提出的梯度提升树结合逻辑回归的方法（GBDT+LR）[3]。

（2）.深度学习模型为卷积神经网络（CNN）。

另外，我也尝试了阿里妈妈的 Multiple Logistic Regression（MLR）[4]，但是在实际运行过程中，我仔细研究了论文和 demo 代码，在本例中的表现欠佳，因而不做过多介绍，而放在附件中。接下来我将对他们进行简单介绍，并附上为什么我会使用他们

1 逻辑回归

在高潜预测中,逻辑回归能够考虑到用户、场景、上下文等丰富的特征进行综合的预测，生成较为全面的推理结果。逻辑回归因其直观简单，可解释性强，运行效率高，易并行等优点，也在业界中作为基础模型或最终输出层模型被广为应用。

同时，逻辑回归也作为感知机的基础作为神经网络中最为基础的单一神经元，是深度学习的基础型结构。在深度学习兴起之前，很多 CTR 模型在逻辑回归的基础之上，进行多特征融合，而衍生出了非常多有意思的模型，如下文中要提到的 GBDT+LR，和未能完满实现的 MLR，还有 POLY2 和 Factorization Machine 等模型。

因此,考虑到其简单,易实现,且具有理论基础意义而作为 Baseline 模型加入本项目中。

2 GBDT + LR

逻辑回归本身对于特征的选择停留在特征本身，而不会考虑到特征之间的交叉，而 FM，POLY2 等模型（对特征进行进一步交叉，如两两配对、三三配对，再喂进逻辑回归）则容易出现维数爆炸的问题，较为消耗算力和时间，且特征处理同预测目标之间的关联性较小。

此时，GBDT+LR 的组合模型解决方案就能够比较有效地处理高纬度特征组合和筛选的问题。模型通过 GBDT 进行特征组合筛选，并将样本置于 GBDT 各个子树的的叶子节点的位置作为输入，喂给逻辑回归，不仅能够减少人工筛选特征和模型设计的精力，更能够实现朝着优化预测结果为目标进行特征的筛选和组合。

GBDT 解决的，是一棵树的表达能力很弱，不足以表达多个有区分性的特征组合，多棵树的表达能力更强一些，且随着模型训练的轮次增加，会通过新的子树来减少同许纳林目标之间的差距，可以更好的发现有效的特征和特征组合。因此在不断分裂的过程中，能够发现对当前步骤判别更有效的特征组合。

3 卷积神经网络（CNN）

我所理解的卷积神经网络，在这里的二分类问题中能够从部分特征在更高维度组合（卷积层映射）之下，找到其中的隐含规律。最终结果会通过类似逻辑回归的形式进行判别，输

[3] He, X. , Bowers, S. , Candela, J. Q. , Pan, J. , Jin, O. , & Xu, T. , et al. (2014). Practical Lessons from Predicting Clicks on Ads at Facebook. ACM.

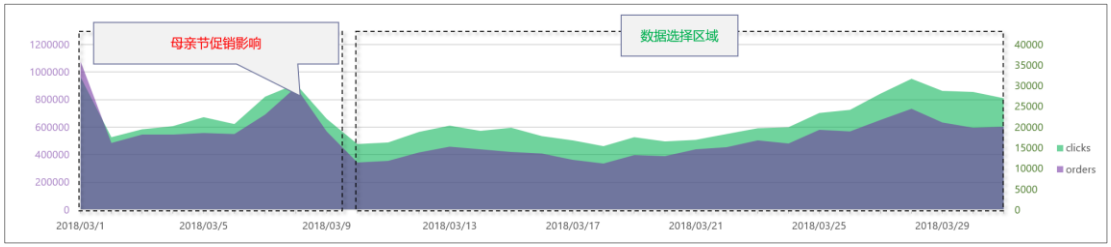
[4] Gai K , Zhu X , Li H , et al. Learning Piece-wise Linear Models from Large Scale Data for Ad Click Prediction[J]. 2017.

出对应类型的概率。因此在构建过程中，在输入层上，我选择以训练、测试数据维度作为输入层，随后进行卷积，并最终利用 sigmoid 完成输出。

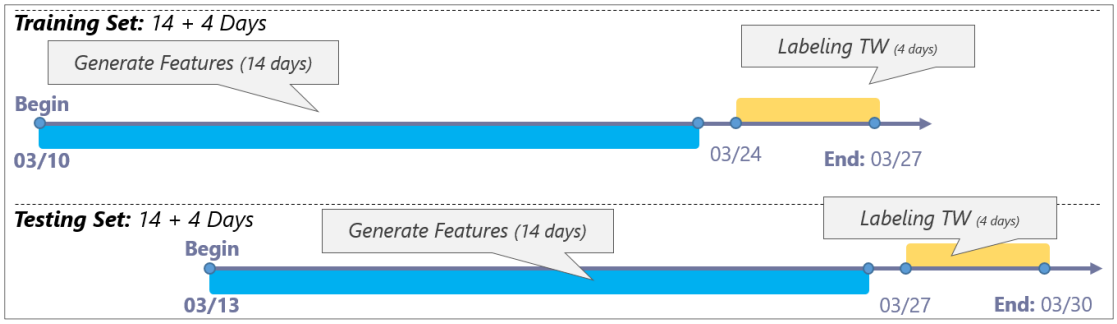
三、实验

第一节 训练集与测试集构建

1 时间窗口选择



首先，收到母亲节促销影响，此品类商品的访问和购买达到了一个峰值，为了避免这里数据的影响，我选择 3 月 10 至 31 日作为我构建训练、测试集的窗口。



我利用 3 月 10 日至 3 月 27 日作为我的训练集时间窗口，其中 3 月 10 日至 23 日提取用户特征，统计用户行为画像，并以此为 X 来预测其在 24 至 27 日（接下来的 4 天）内是否会购买。

同理处理测试集（3 月 13 日至 30 日）。

2 入模特征

根据题目所提供的特征，我分为了这样几种类型和对应的处理方法：

类型	字段	含义	来源
布尔型	plus	是否为 plus 用户	user.plus
	is_new_user	是否为新用户用户	user.purchase_power
	is_enterprise	是否为企业用户	user.purchase_power
离散型	age	年龄区间	user.age

	gender	性别	user.gender
	marital_status	婚姻状态	user.marital_status
连续型	education	教育程度	user.city_level
	city_level	城市消费力	-
	purchase_power	推定用户消费水平	-
	jd_age	京东用龄	user.first_order_month
	clk_1d	-	行为画像
	clk_3d	-	行为画像
	clk_7d	-	行为画像
	clk_14d	-	行为画像
	clk_distract_score	-	行为画像
	odr_1d	-	行为画像
	odr_3d	-	行为画像
	odr_7d	-	行为画像
	odr_14d	-	行为画像
	odr_distract_score	-	行为画像

- 对于布尔型特征不需要进行额外处理；
- 对于离散性特征，通过转换成 One-Hot 编码。这里的考量是：题目给定的这部分字段已经是预测出来的字段了，此时尚有位置的情况，不想贸然用模型进行缺失值填充；
- 对于连续型特征，通过 StandardScalar 方法进行 z-score 标准化处理；

第二节 Metrics 选择

从业务角度，需要关注的目标为：模型是否能够尽可能找到多的高潜用户，和模型是否足够强健。

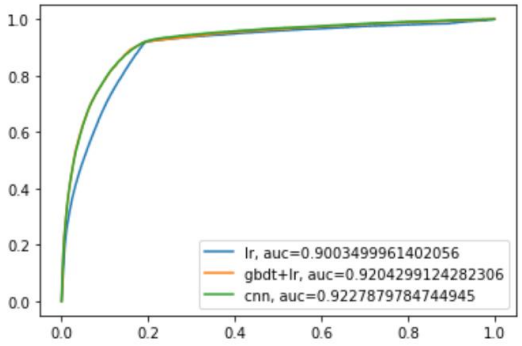
基于此，我选择对正类的 precision，以及模型的 AUC 作为衡量指标。

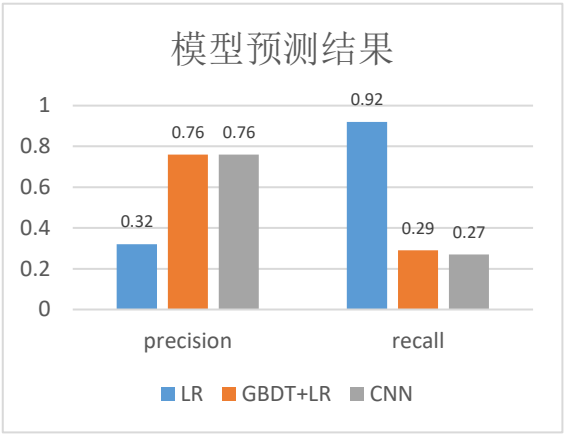
第三节 模型实践

（详见 ipynb 代码）

四、分析与结论

第一节 效果分析





从 AUC 曲线上看，CNN 的表现最好，随后是 GBDT+LR，最后则为 LR，可以看出他们三者的对正例的识别能力。

与此同时，从 precision 和 recall 的角度上看，逻辑回归作为 baseline 和其他二者相比，虽然 recall 更高，但是 precision 却低了很多。

这是由于样本的不平衡所导致的，在本例中，正负例比约为 18：1，而逻辑回归尽管能够找到绝大部分的正例，具有较高的召回值，但其误判了很多负例为正（假阳性过多），因而在查准率上得分较低。而后两者明显更能够在不平衡的情形下找到正例更显著的特征，这和先前提到的模型逻辑中构建的思路相吻合——GBDT 在特征组合之间不断逼近真相，而 CNN 在高维映射中找到特征组合之间的特点。

五、总结与寄语

第一节 总结

通过复现经典的模型完成了高潜用户的预测，尝试从模型的逻辑和栈角度出发，在案例中实践学习，加深了我对他们的认识。

第二节 展望

1 数据部分

在特征部分，我仅仅将他们进行粗略的统计和转换用户基本属性（人口统计学特征、推断特征、行为画像），而没有考虑到其他用户消费场景当中的因素：如用户对商品访问序列、购买序列，从这个角度去刻画用户在消费角度上的偏好。

这是由于我不是很清楚对于此类序列型数据应该如何处理和表现。尽管有 embedding 方法能够利用类似自然语言的方法对行为序列进行编码和表征，但是并不能区分出点击、购买的差异，以及不同时间下的访问差异。

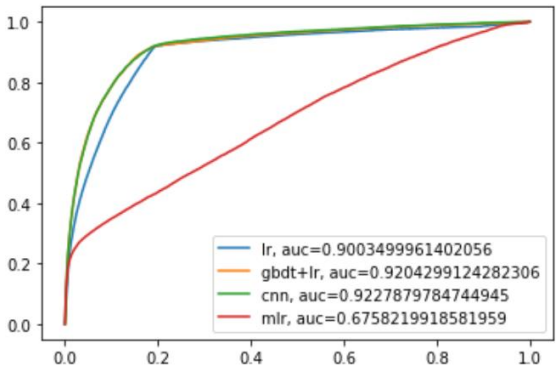
出于此，我简单探索了 Item2Vector 的方法，对 sku 进行了处理和编码（放在附件中），但未对用户访问 sku 的序列进行操作。

出于常识，在判断高潜的过程中，除开单纯的统计型数据，如果更能够描述用户的浏览信息，相比更能够有价值和意义。

因此，我认为在此基础之上，可以考虑对样本进一步进行区分，并研究一下深度学习在序列中的处理方法，或者加入 LSTM 和 Attention 机制等，尝试找突破口。

2 模型部分（MLR）

由于我对 TensorFlow 的理解较浅，在学习了原论文和作者公布的代码之后，尝试应用在本案例的过程中，出现了异常情况：



随着训练轮次的提高，并没有看到训练、测试数据集上的 auc 有明显提升（详见 notebook），这与作者给出的数据集所体现的效果相反，因而我并将它放在实验中的模型部分，希望能够待到日后，再展开研究。

六、参考资料

[1]. He, X. , Bowers, S. , Candela, J. Q. , Pan, J. , Jin, O. , & Xu, T. , et al. (2014). Practical Lessons from Predicting Clicks on Ads at Facebook. ACM.

[2]. Gai K , Zhu X , Li H , et al. Learning Piece-wise Linear Models from Large Scale Data for Ad Click Prediction[J]. 2017.

[3]. 阿里文娱技术，阿里文娱智能营销增益模型（Uplift Model）技术实践，2020-03-25

[4]. 赵宏田，用户画像——方法论与工程化解决方案，机械工业出版社

[5]. 项亮，推荐系统实践，人民邮电出版社

[6]. 王喆，深度学习推荐系统，电子工业出版社

[7]. 周志华，机器学习，清华大学出版社