AiCon
全球人工智能与机器学习技术大会

商业智能技术驱动业务增长

张重阳

微信小程序商业技术负责人

主办方 Geekbang. 极客邦科技 InfoQ

极客时间

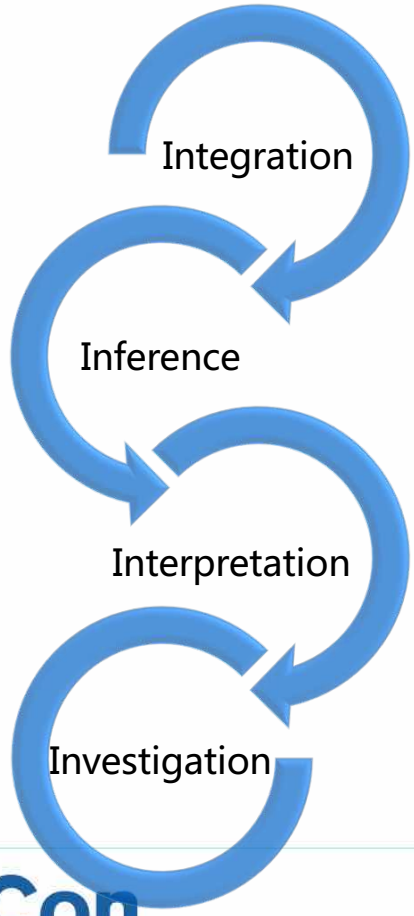
重拾极客精神·提升技术认知

下载极客时间App

获取有声IT新闻、技术产品专栏，每日更新

扫一扫下载极客时间App

智能技术在商业过程中应用的4个环节



- 整合: 针对问题收集并整合数据
  - 如何在计算机中表示数据便于存储和计算
  - 如何处理保密数据，如何在处理用户数据时保护用户隐私
- 推断: 使用统计和机器学习的方法求解问题的最优解
  - 如何结合多个模型优点
- 解释: 结合数据对模型的推断结果进行分析和解释
  - 如何对黑盒模型的结果进行解释
- 调查: 利用人工知识制定策略并使用自动化的方法验证效果
  - 如何在实际环境下验证效果

微信商业智能系统

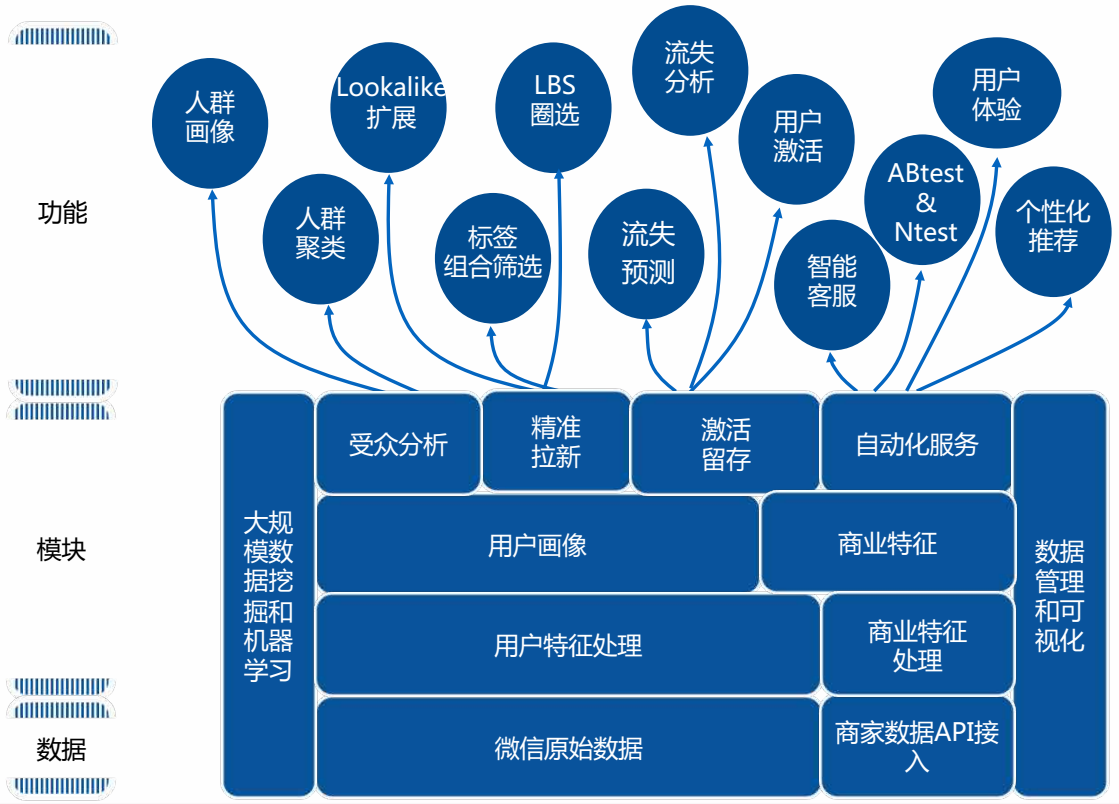

功能
模块
数据
人群画像
人群聚类
Lookalike扩展
标签组合筛选
LBS圈选
流失预测
流失分析
用户激活
智能客服
ABtest & Ntest
用户体验
个性化推荐
受众分析
精准拉新
激活留存
自动化服务
大规模数据挖掘和机器学习
用户画像
商业特征
用户特征处理
商业特征处理
微信原始数据
商家数据API接入
数据管理和可视化

TABLE OF CONTENTES

整合 Integration

- 实例: 用户画像

推断

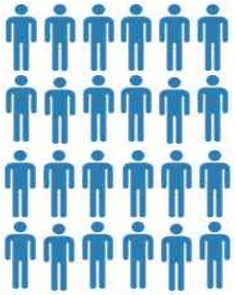
解释

调查

总结

用户画像

用户画像



存储结构

	特征a	特征b	特征c	特征d	…
用户1					
用户2					
…					
用户n					

距离度量

测量用户之间的相似度

应用场景:

- 用户分类，聚类，对相似用户推荐相同商品

用户A: x　用户B: y

相似度函数: F(x, y)

- 实际使用中常根据不同的应用场景定义不同的相似度函数或是用机器学习的方法在数据上拟合这个函数

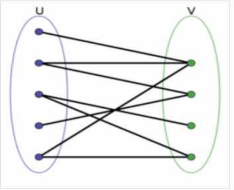
多源异构

可标签化

异构信息

社交网络

公众号关注



短期兴趣

数据编码

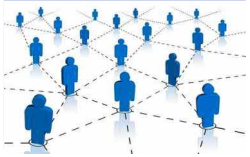
基础数据

- 性别
- 年龄
- 城市
- ...

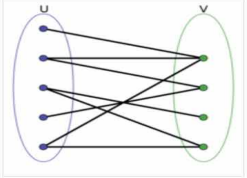
兴趣标签

- 健身
- 理财
- 旅游
- ...

社交网络



公众号关注



短期兴趣

其他

- 文本
- 图像
- ...

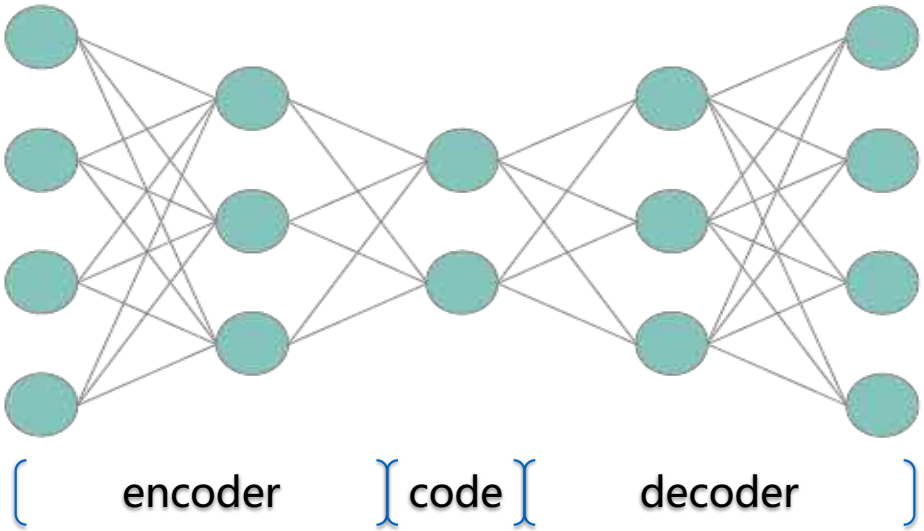
标签	标签	node2vec	Topic Model	Sequence Embedding	encoder
向量	向量	向量	向量	向量	向量

用户特征向量

降维

Auto-encoder

- 降维可以使实际应用的机器学习任务支持更小的样本输入
- 降维后的特征向量对用户之间的距离度量影响不大，但相当于原始特征能更好的保护用户隐私



用户隐私保护遵守的原则

1. 分析一群人而不分析一个人。
2. 不使用个人可辨识信息（Personal Identifiable Information），如：姓名，身份证号，手机号 等。（我们数据处理时使用无任何物理含义的User ID作为各个数据中的统一标识）
3. 通讯和聊天内容神圣不可侵犯，不保存和使用任何通讯和聊天内容。
4. 控制精度，这里的精度并不是指准确度，比如我们在分析用户住址时，只定位到小区，而不再做楼栋和楼层的定位。
5. 只保留和使用一年以内的数据。
6. 所有的标签都由算法自动化生成而不使用人工标注，工程师只负责设计算法。

TABLE OF CONTENTES

整合

推断 Inference

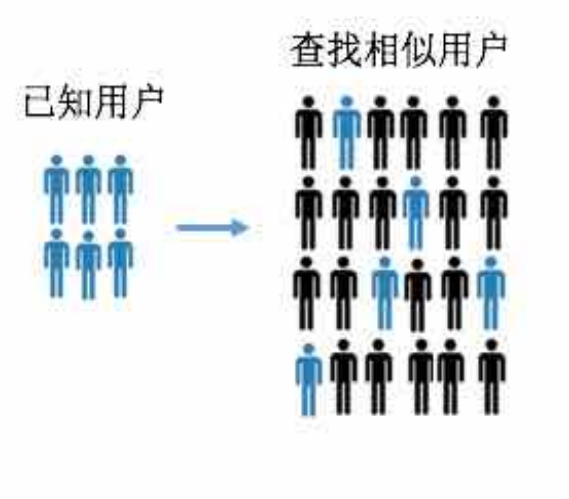
- 实例: Lookalike人群定向与流失预测

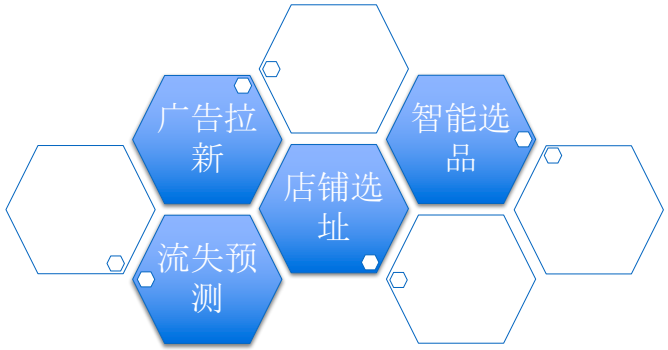
解释

调查

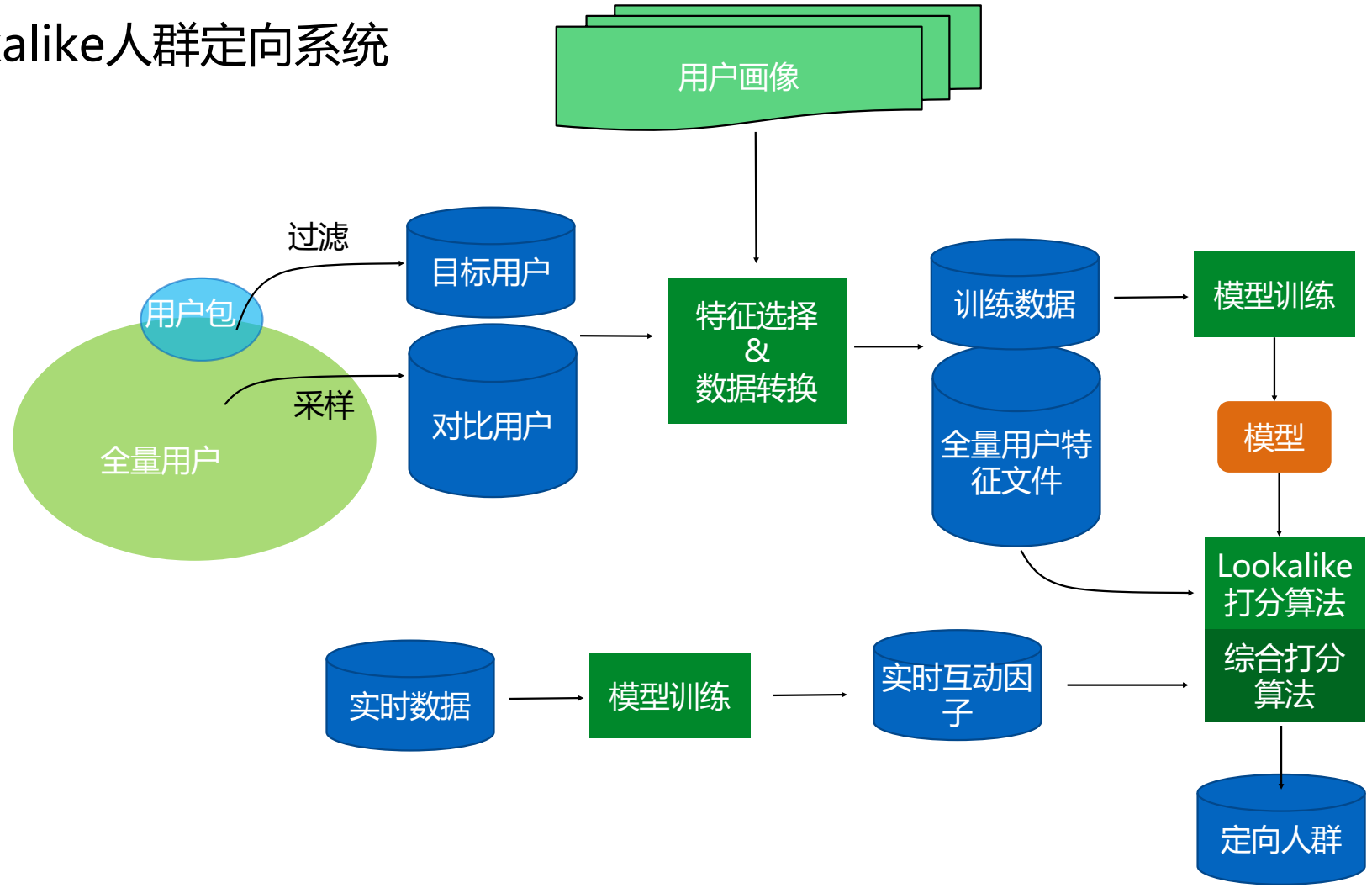
总结

Lookalike潜在应用场景



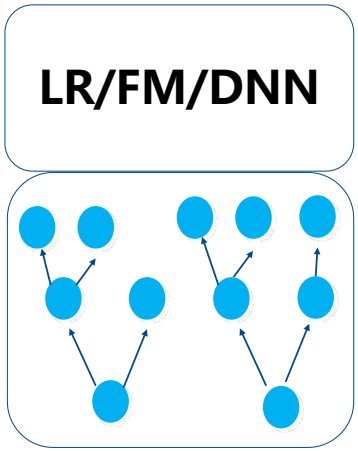


Lookalike人群定向系统



模型整合 (JointTrain vs Ensemble)

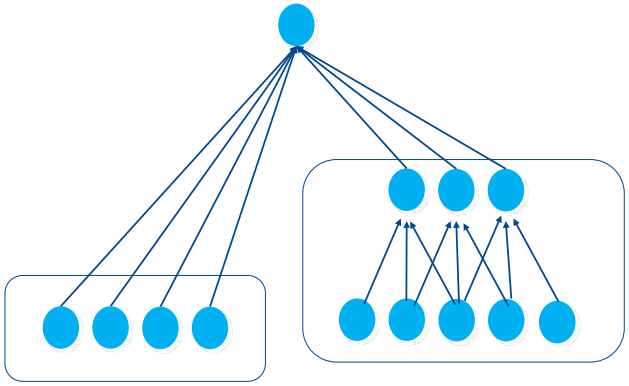
集成学习



GBDT2LR
GBDT2FM
GBDT2DNN

- 通用方法

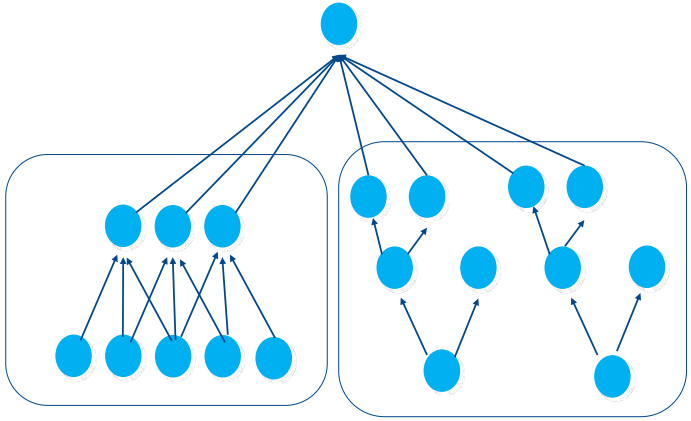
联合训练



Wide & Deep

- 可以结合人工定义的特征

集成学习



DNN + GBDT

我们系统实际采用的模型

- 不需要加入人工特征，且实际效果更好

应用实例：流失预测与流失分析

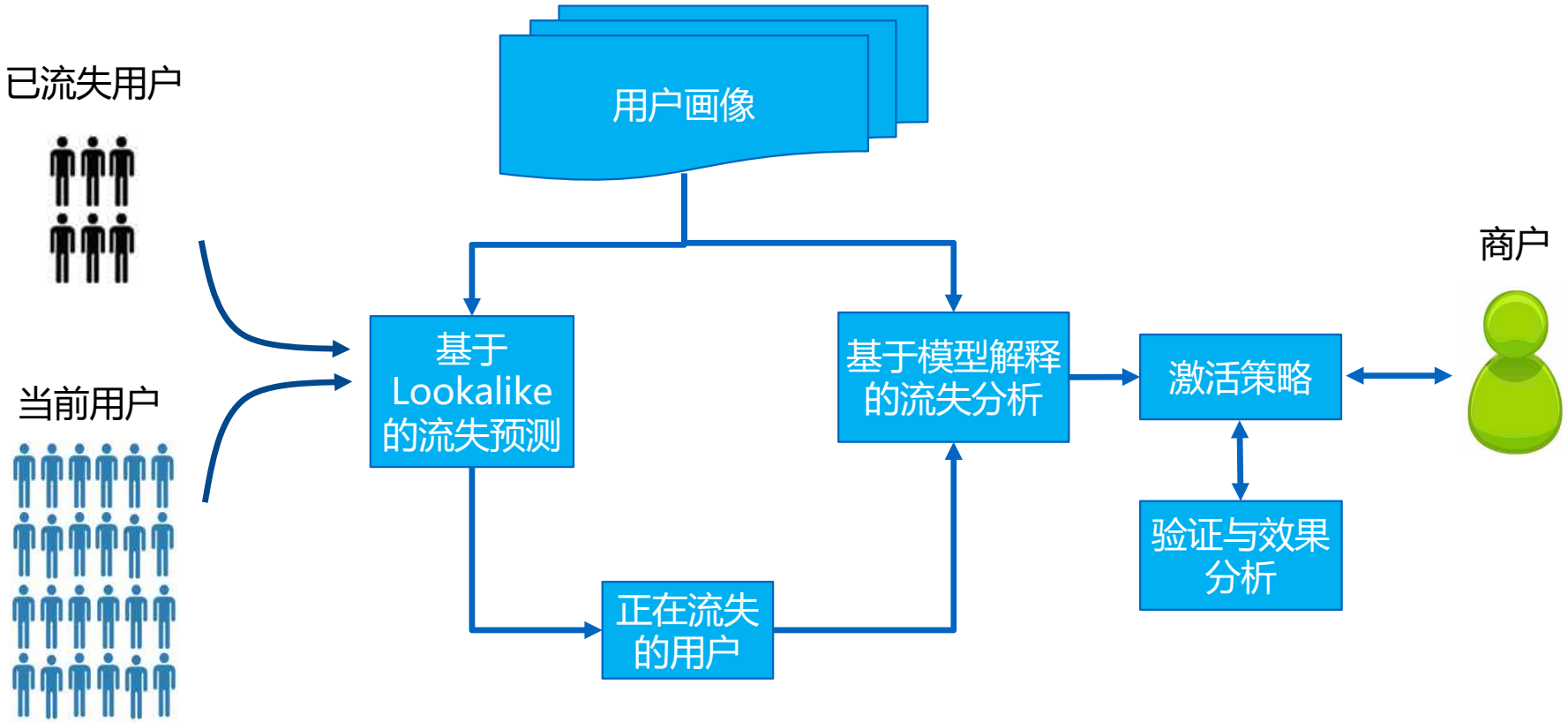


TABLE OF CONTENTES

整合

推断

解释 Interpretation

- 实例：流失分析

调查

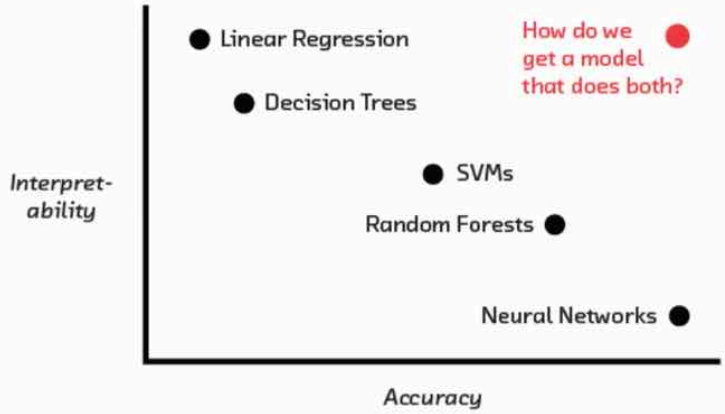
总结

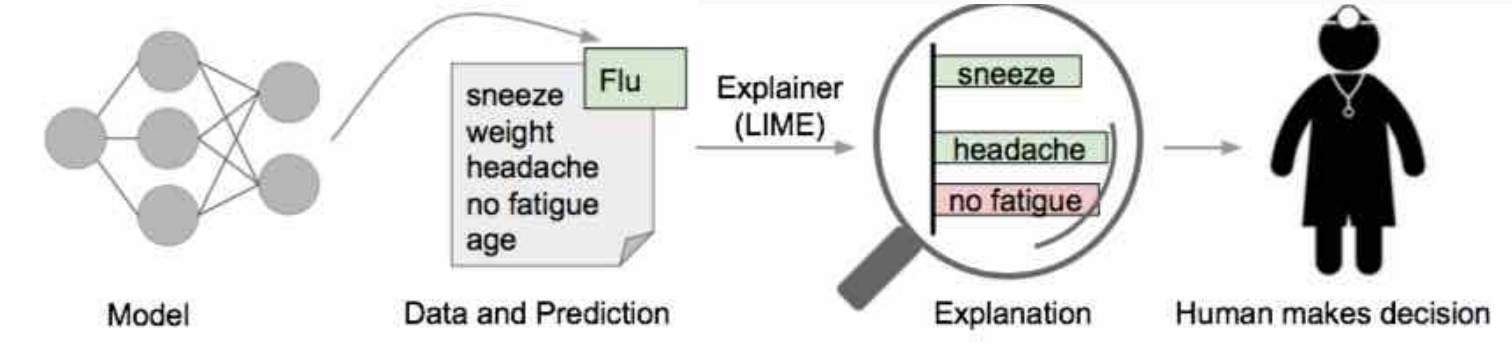
模型解释

模型的可解释性和精度同等重要，然而模型的能力越强就越复杂越不易解释

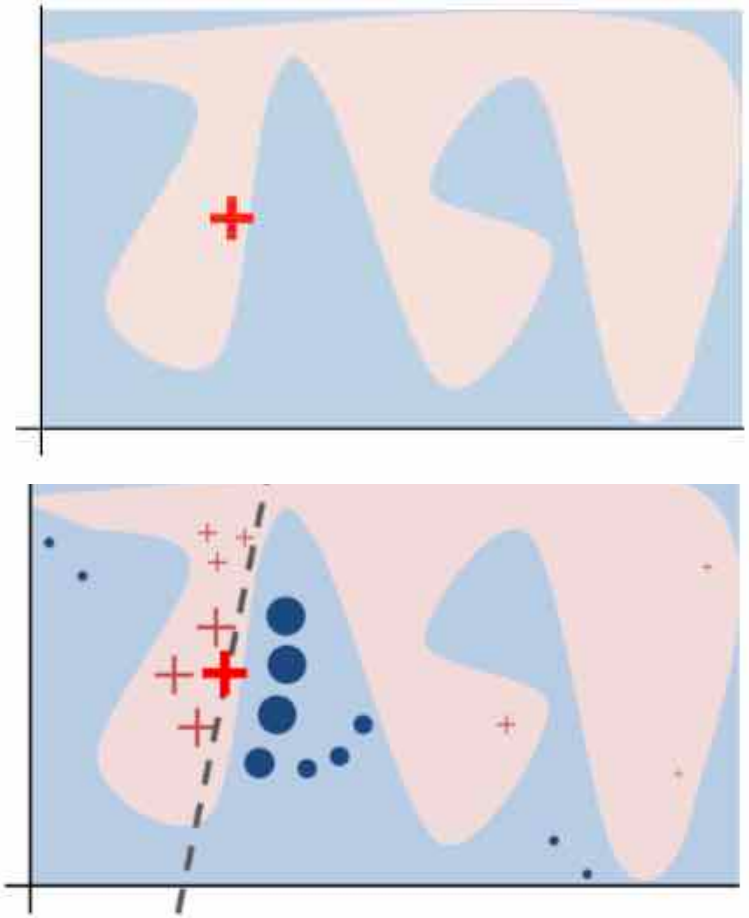
对模型的有效解释可以：

- 发现训练数据和模型中的问题
- 更容易说服模型输出结果的反馈对象
- 从解释结果中获取更多信息从而制定策略解决问题





模型无关的局部解释算法LIME

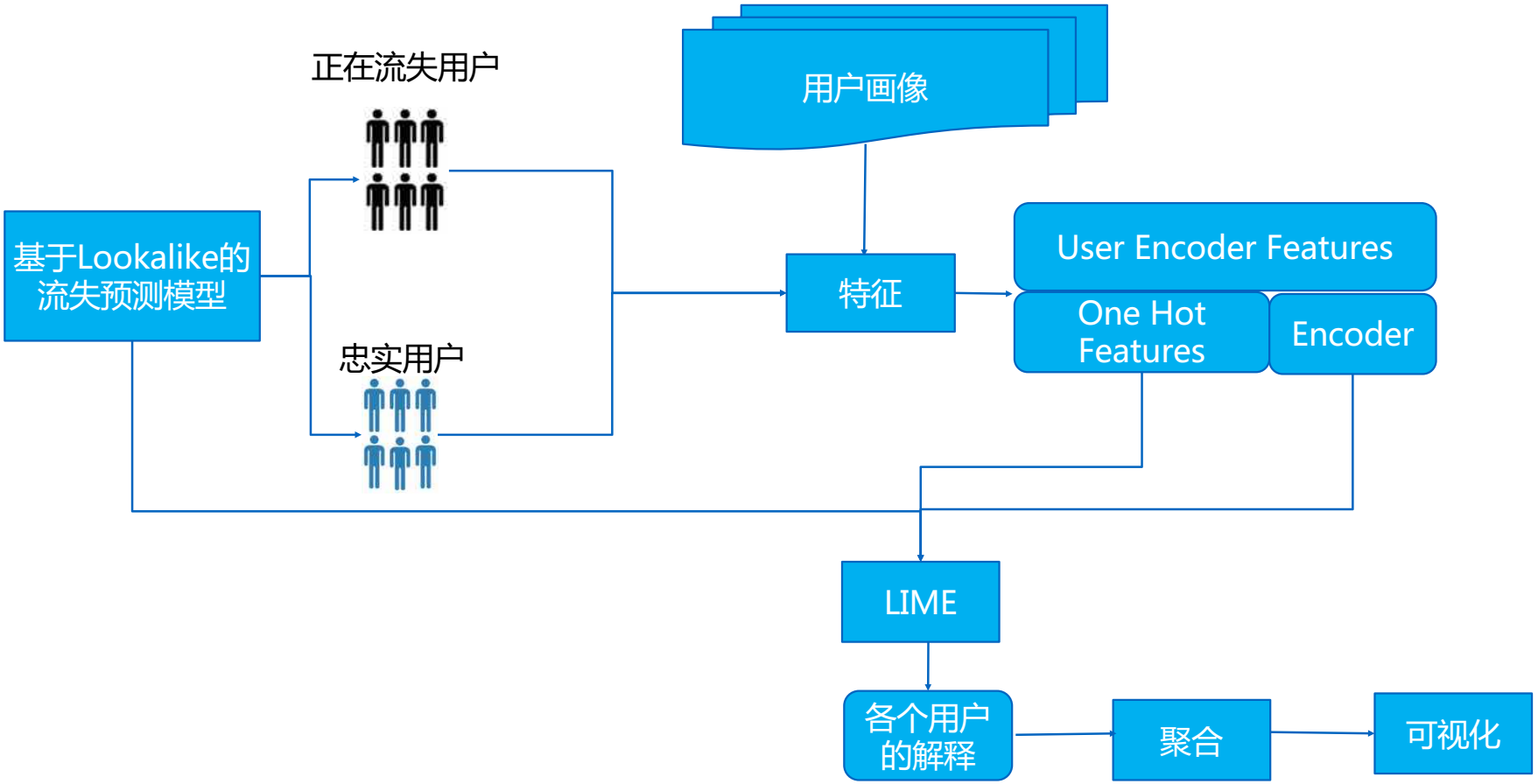


KDD 2016： “Why Should I Trust You?” Explaining the Predictions of Any Classifier

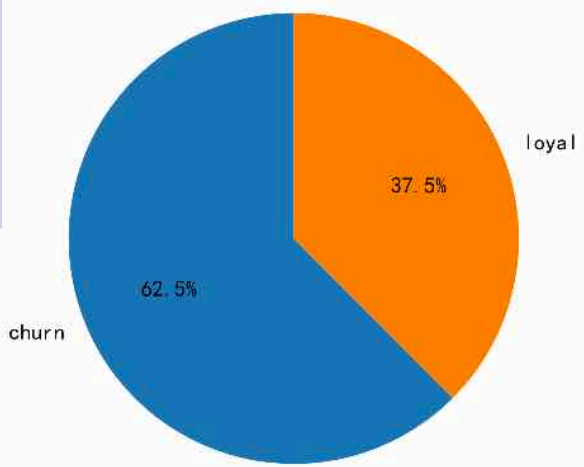
Algorithm 1 Sparse Linear Explanations using LIME

Require: Classifier f, Number of samples N
Require: Instance x, and its interpretable version x'
Require: Similarity kernel π_x, Length of explanation K
  $\mathcal{Z} \leftarrow \{\}$
  for $i \in \{1, 2, 3, ..., N\}$ **do**
    $z'_i \leftarrow sample\_around(x')$
    $\mathcal{Z} \leftarrow \mathcal{Z} \cup \langle z'_i, f(z_i), \pi_x(z_i) \rangle$
  end for
  $w \leftarrow$ K-Lasso$(\mathcal{Z}, K)$ ▷ with z'_i as features, $f(z)$ as target
  return w

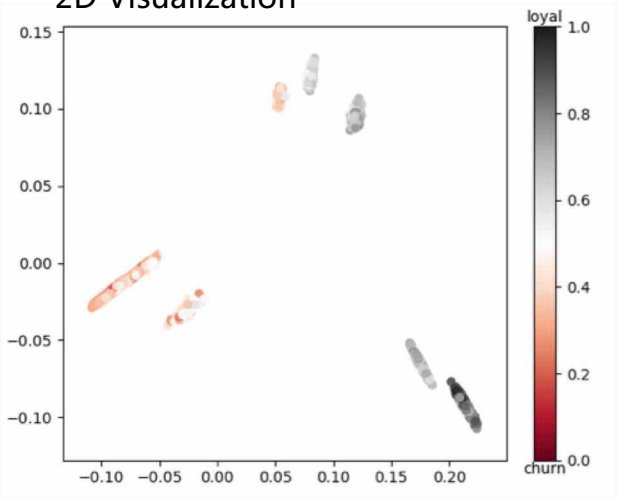
应用实例：用户流失分析

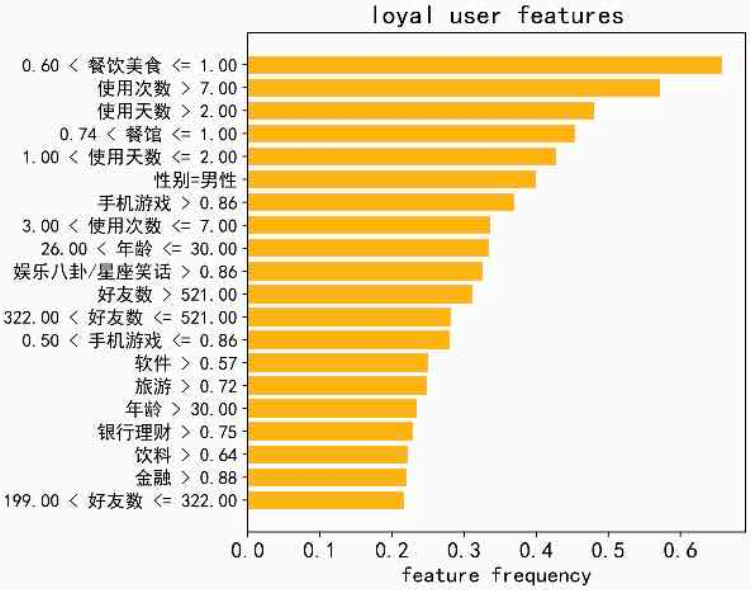

正在流失用户
用户画像
基于Lookalike的流失预测模型
特征
User Encoder Features
One Hot Features
Encoder
忠实用户
LIME
各个用户的解释
聚合
可视化

某小程序用户流失分析结果
(流失预测AUC为0.8)



2D Visualization





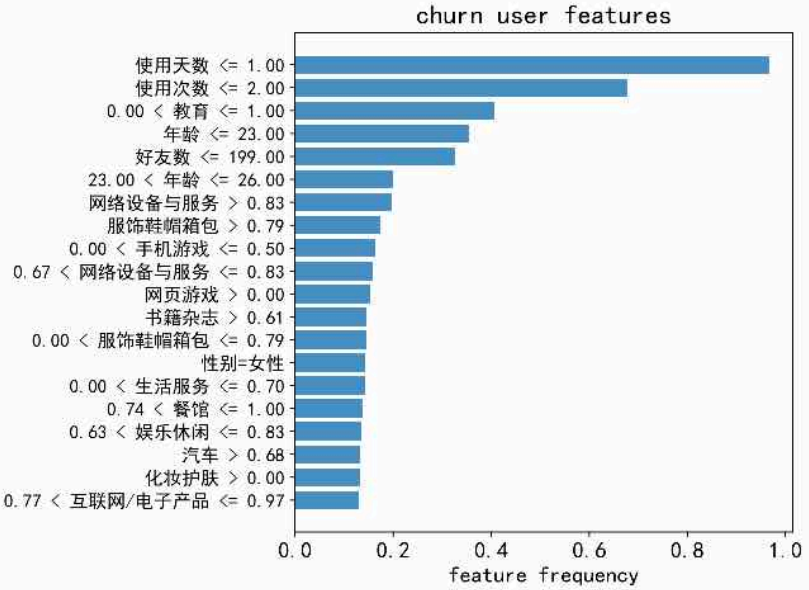


TABLE OF CONTENTES

整合

推断

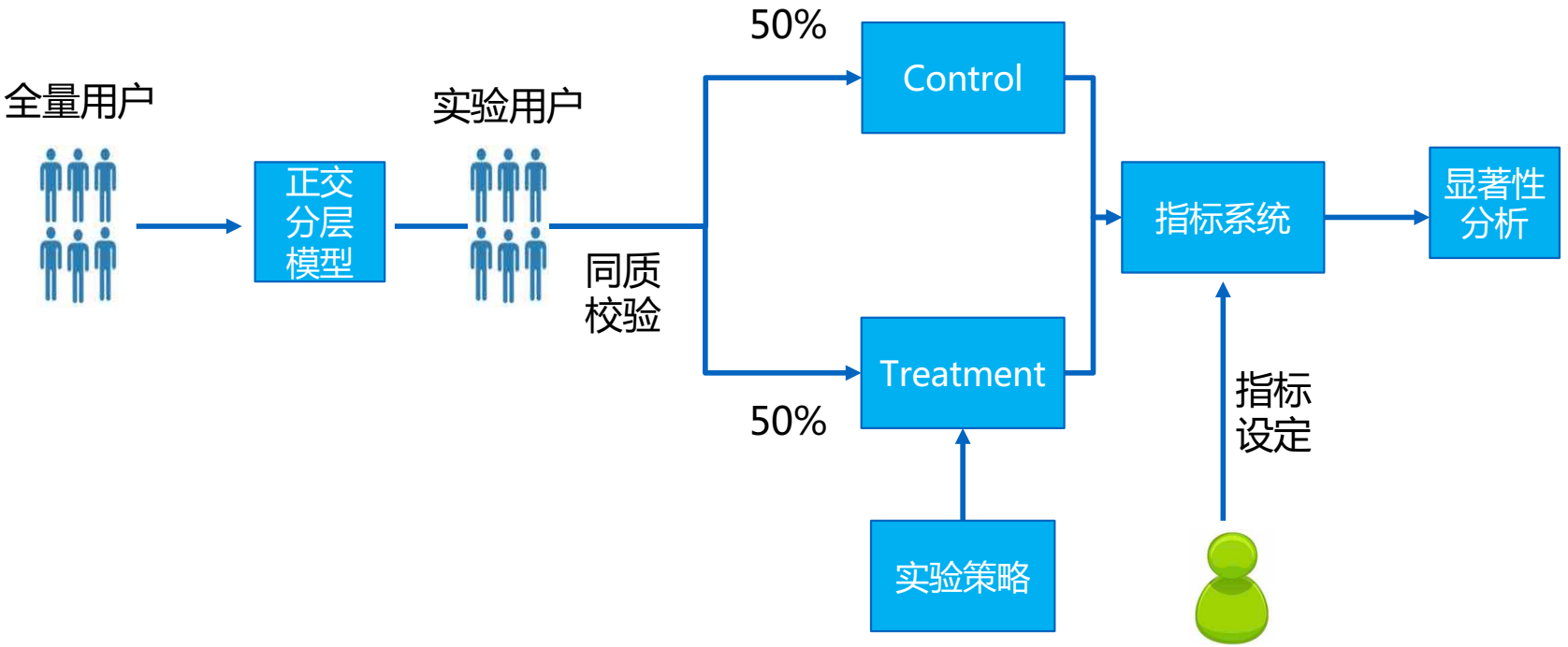
解释

调查 Investigation

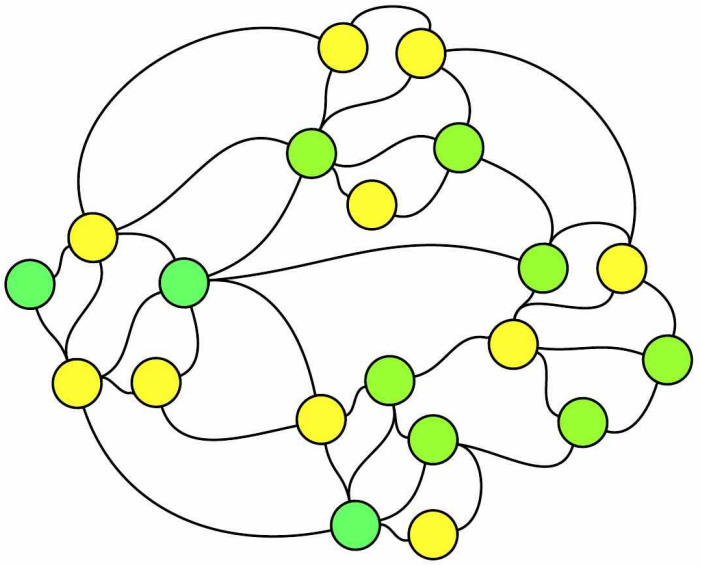
- 实例: Ntest

总结

传统ABtest系统


全量用户
正交
分层
模型
实验用户
同质
校验
50%
Control
Treatment
50%
实验策略
指标系统
显著性
分析
指标
设定

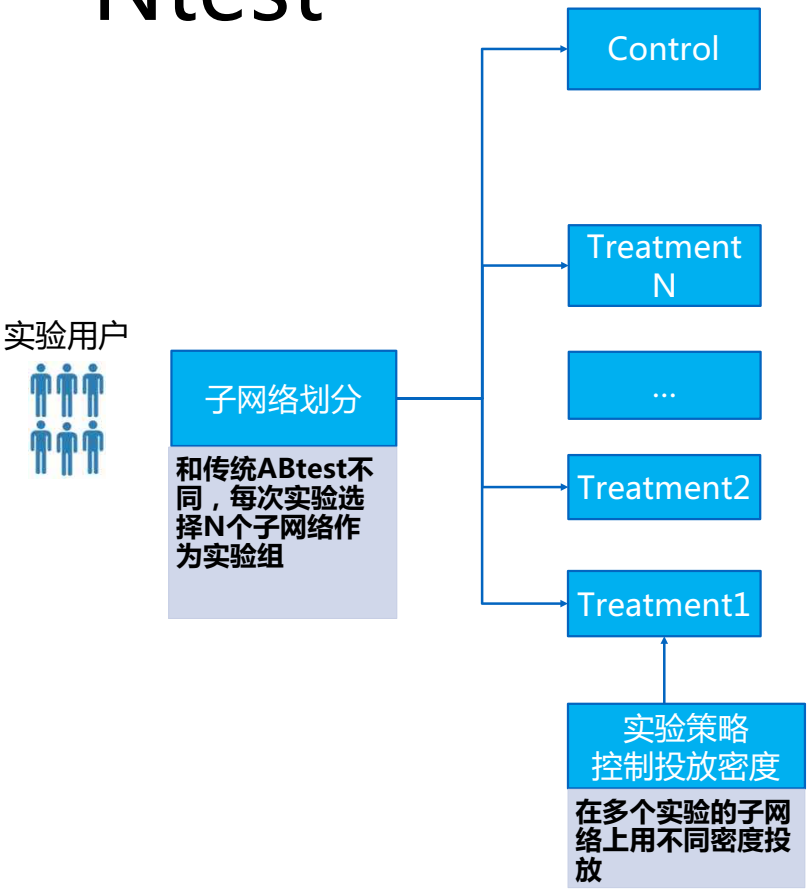
社交网络上ABtest的问题



社交网络上ABtest的问题

- A集合和B集合之间相互影响
- 在小样本的测试下有效，逐步加量后是否会继续有效，如何分析投放密度对效果的影响

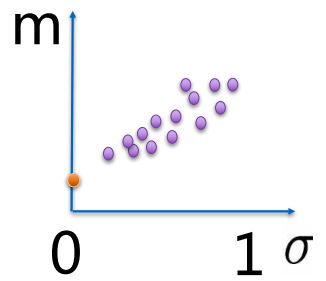
Ntest



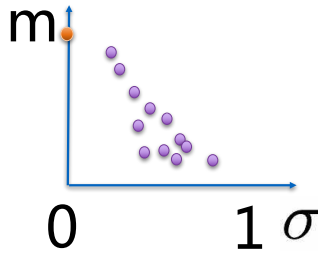
绘制各个指标实验效果和投放密度的关系

m：指标

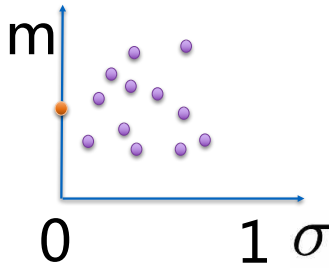
σ：密度



正相关



负相关



无关

TABLE OF CONTENTES

整合

推断

解释

调查

总结

- 人工智能和机器学习落地实践心得

总结：机器学习落地实践心得

Thanks