DataFunSummit

数据科学

在线峰会

金融数据科学 论坛

2021.06.26（周六） 09：00~17：30

2021

度小满金融 | DataFunSummit

数据科学在金融风控模型中的应用

严 澄　　度小满风控模型负责人

目录

CONTENTS

01

科学定义数据

Subject

如何对齐模型目标和业务目标?

金融风险管理

风险匹配

如何定义风险？

如何预测风险？

科学定义数据

$$\text{年化风险} = \frac{\text{年化不良金额}}{\text{年化余额}}$$

业务指标（耦合很多因素，如久期/额度/定价）

$$\text{人数逾期率} = \frac{\text{逾期用户数}}{\text{总用户数}}$$

模型目标(简洁明了)

$$\frac{\text{年化风险}}{\text{人数逾期率}}$$

- > 1：意味着头部用户给的额度过高
- 接近1：额度和风险较匹配
- < 1：意味着尾部用户给的额度较低，制约了整体规模

科学定义数据

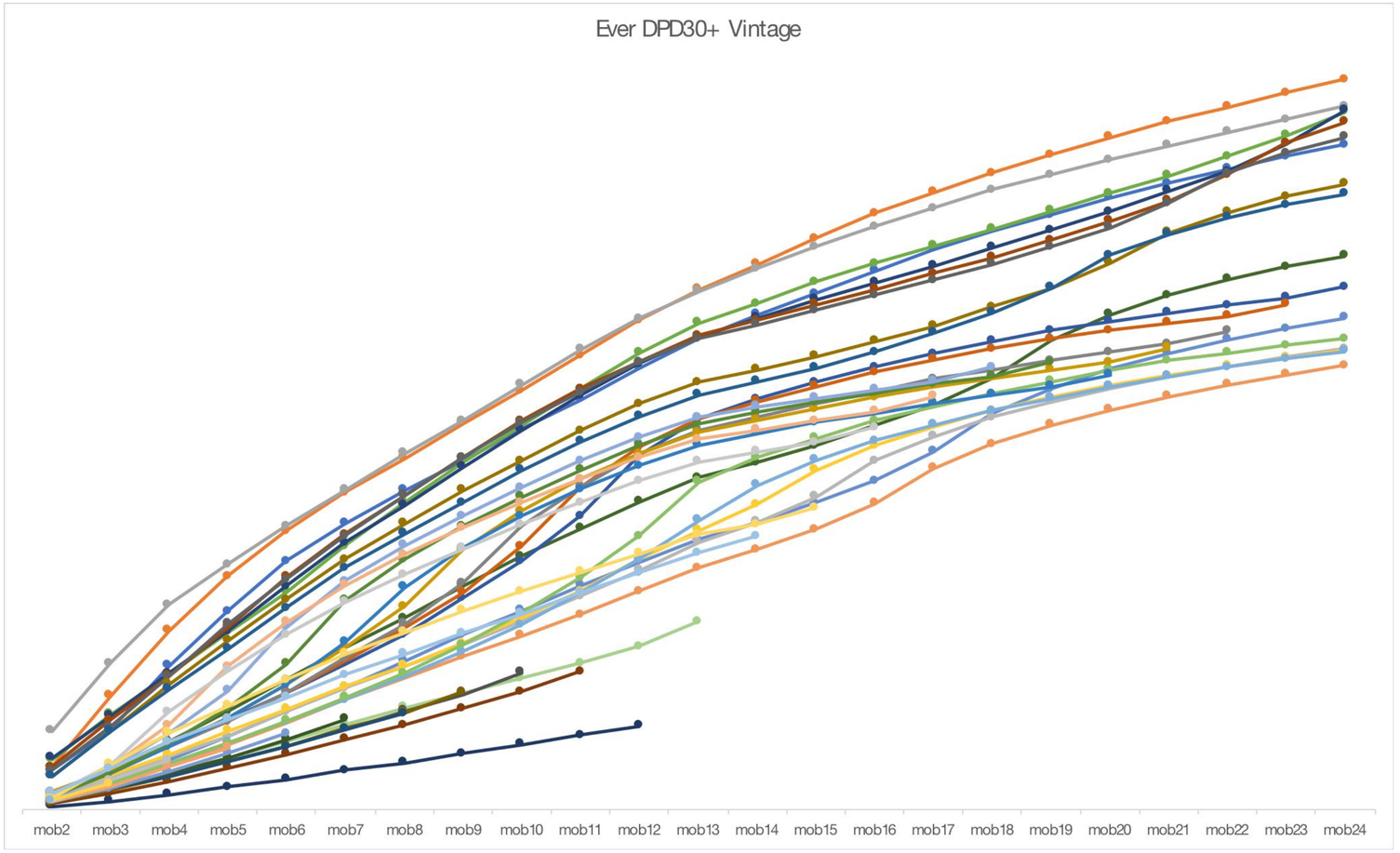
样本和标签选择

*y*如何定义?

- “好”用户：多久不逾期?
- “坏”用户：逾期多久?

观察期越长

- 利：标签置信度高
- 弊：样本量少，且“不新鲜”



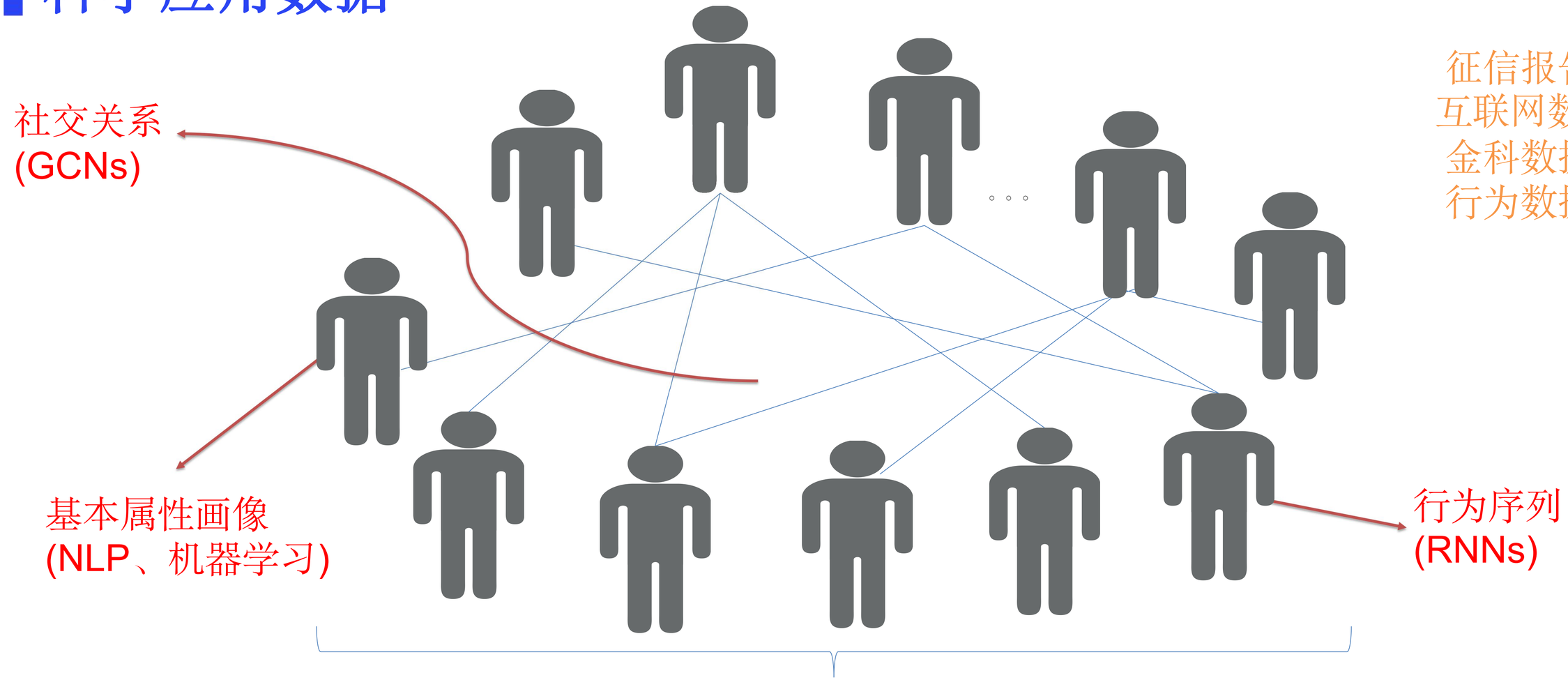
02

科学应用数据

Subject

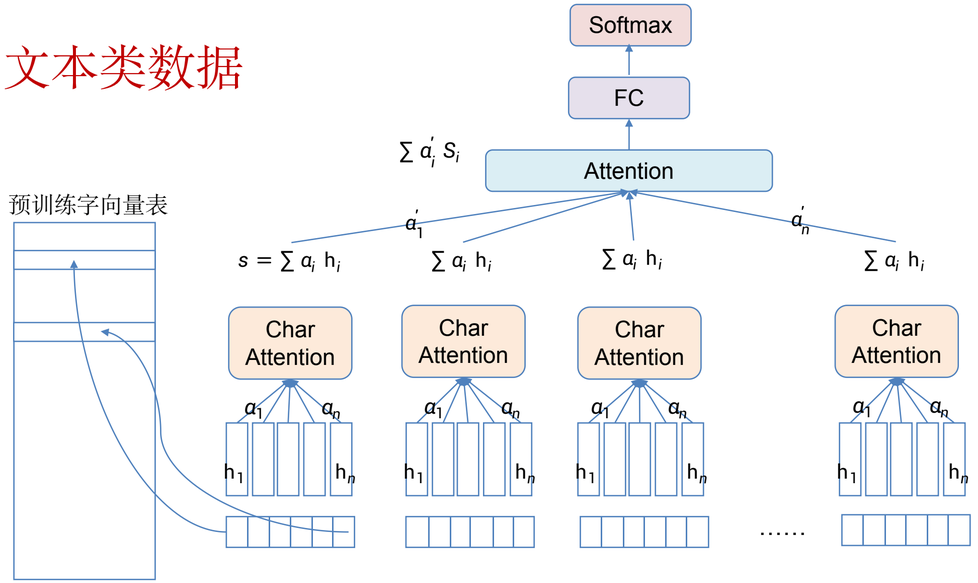
哪些数据可以应用于风控模型？

科学应用数据

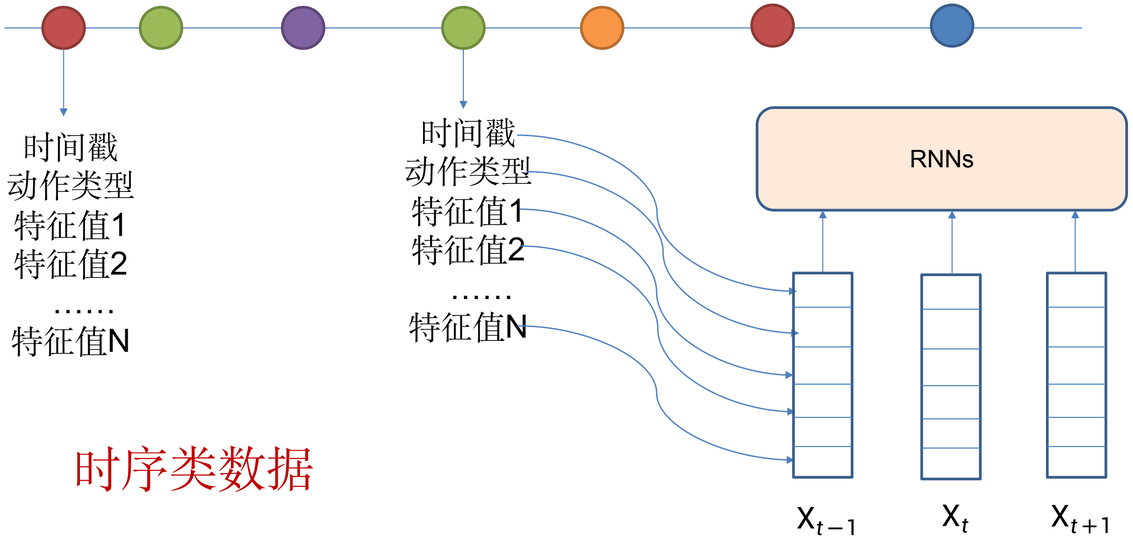

社交关系
(GCNs)
征信报告
互联网数据
金科数据
行为数据
基本属性画像
(NLP、机器学习)
行为序列
(RNNs)
选择合适的统计学习样本

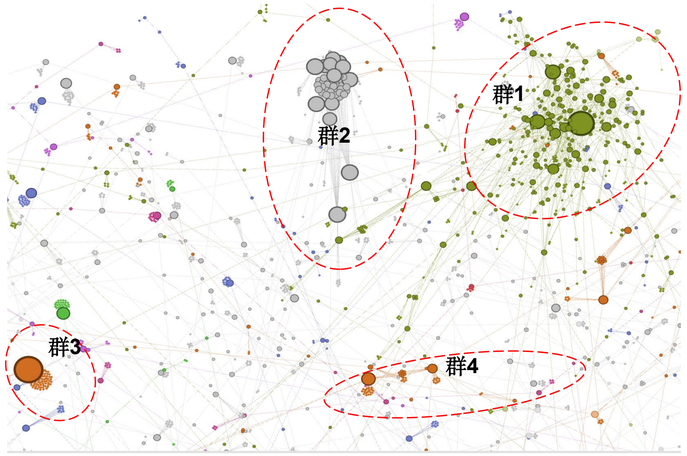
科学应用数据

文本类数据



时序类数据

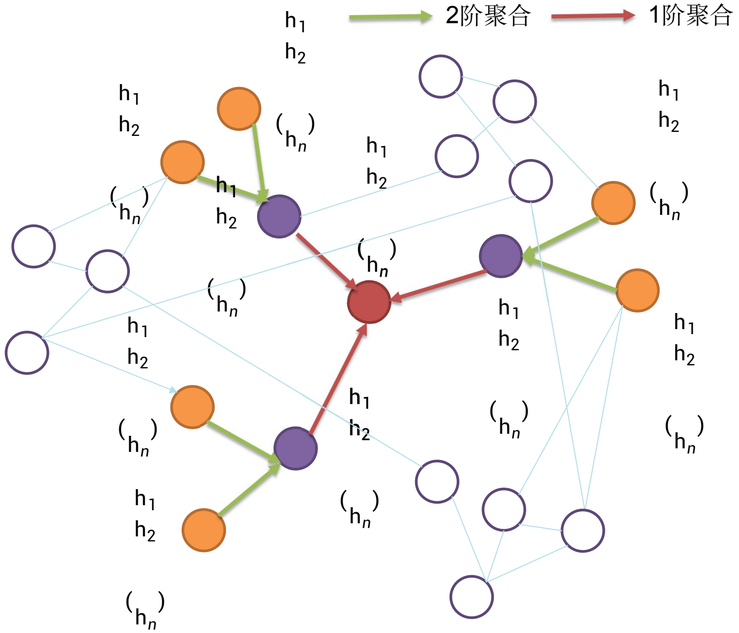




群属性

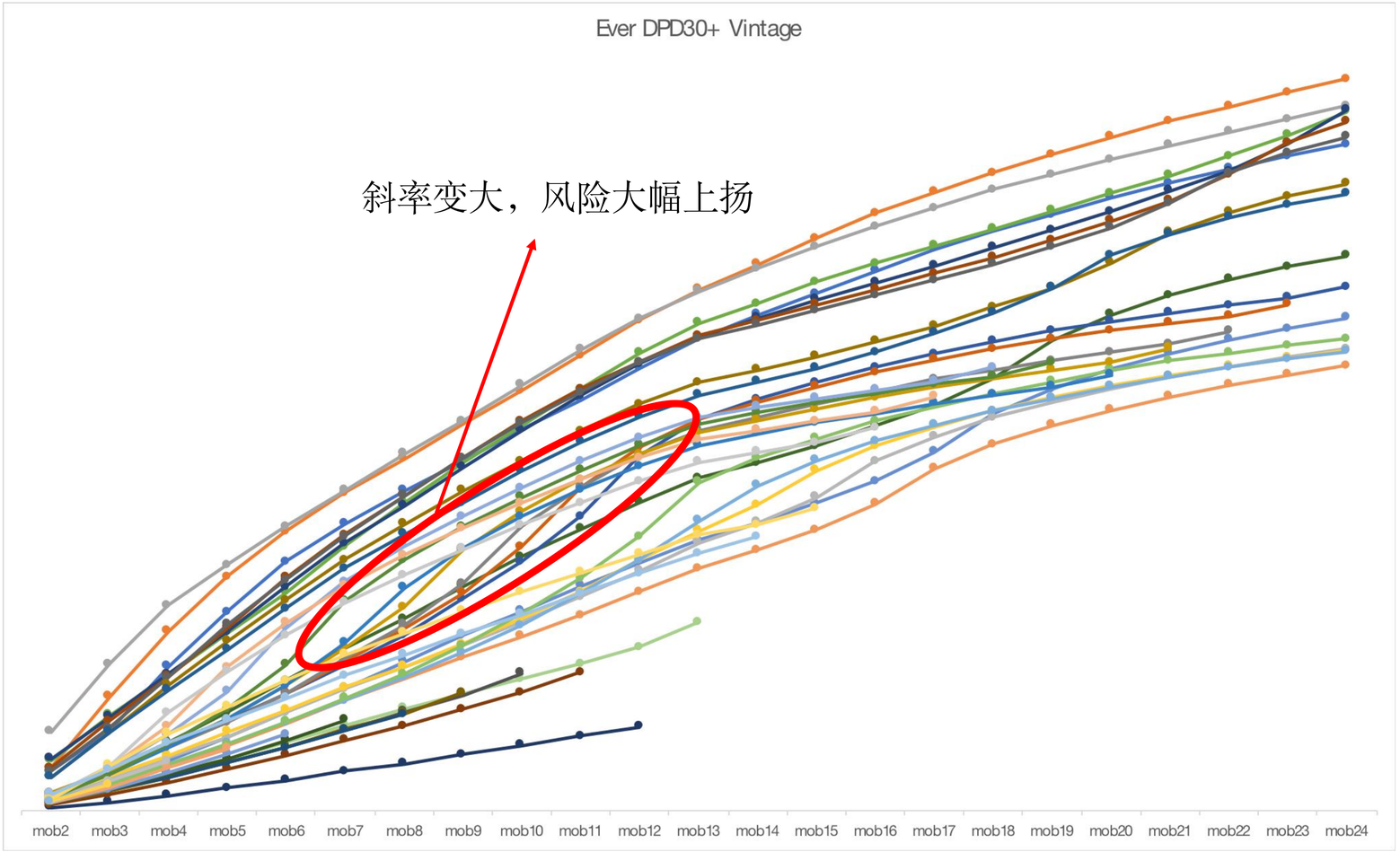
- 群内节点总数
- 群内边总数
- 群内申请用户占比
- 群内用信用户占比
- 群内平均逾期率
- 群内地域分布
- 群内男女分布
- 群内年龄分布
- 群内职业分布
- .....

关联类数据



GCNs

科学应用数据



1. 疫情影响是否结束？

2. 疫情影响了哪些样本？

3. 如何利用疫情下的样本？

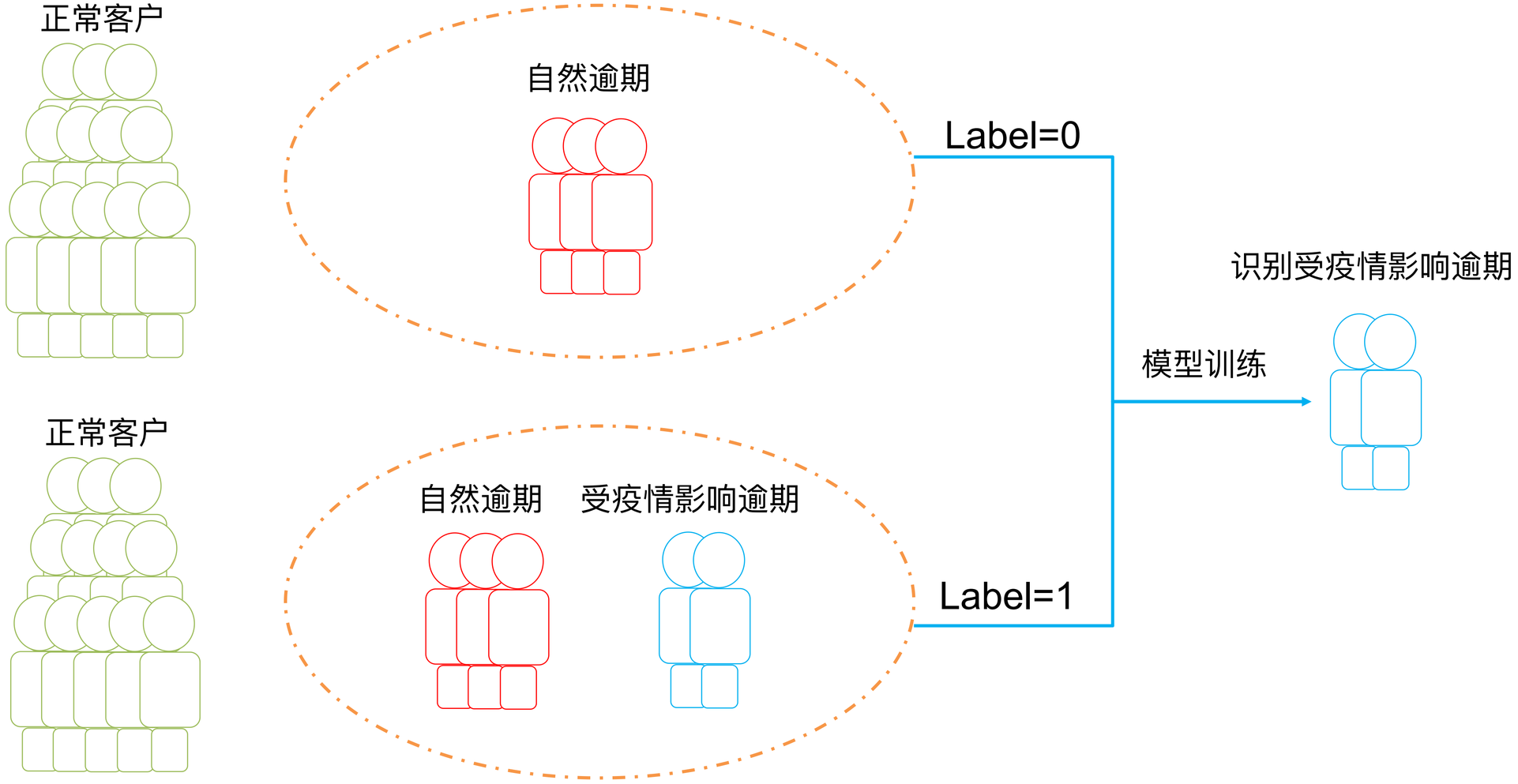
$$p = P(y = 1|X) = f(x_1, x_2, x_3, \ldots, x_n)$$

p ↑　　X →

科学应用数据

表现期不在疫情期间

表现期在受疫情期间

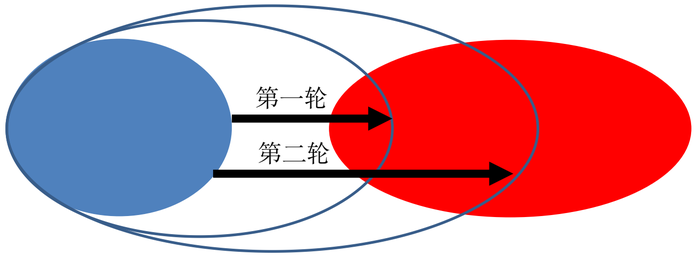


科学应用数据

逾期用户分布

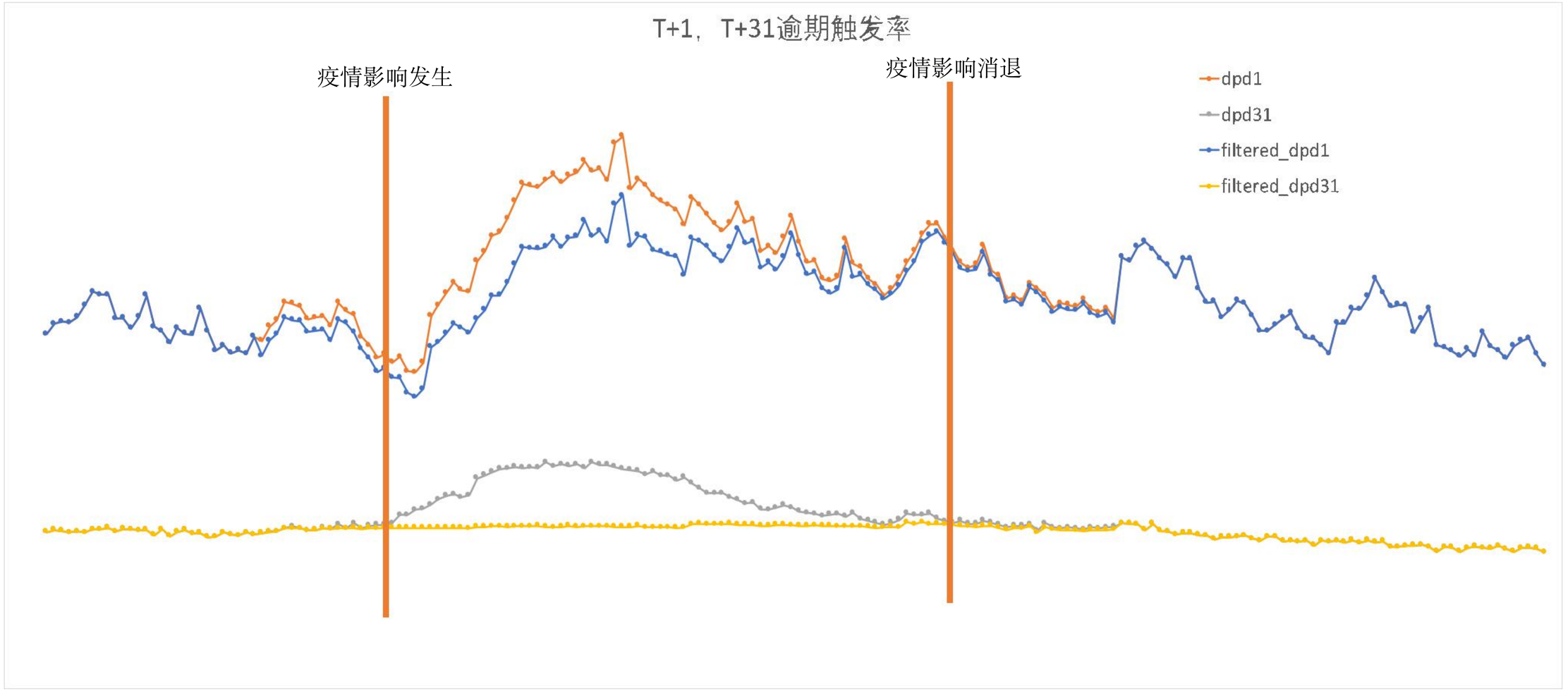
	mob2	mob3	mob4	mob5	mob6	mob7	mob8	mob9	mob10	mob11	mob12
201809											
201810											
201811											
201812											
201901											
201902											
201903											
201904											
201905											
201906											
201907											
201908											
201909											
201910											
201911											
201912											
202001											
202002											
202003											

EM迭代



当风险水平接近疫情前时终止

科学应用数据


T+1，T+31逾期触发率
疫情影响发生
疫情影响消退
dpd1
dpd31
filtered_dpd1
filtered_dpd31

03

科学评估数据

Subject

如何准确地评估模型效果？

科学评估数据

排序性： KS

- 不同评估集上，KS绝对值没有可比性。
- 上线决策后必然衰减，KS提升幅度越大，衰减越大。

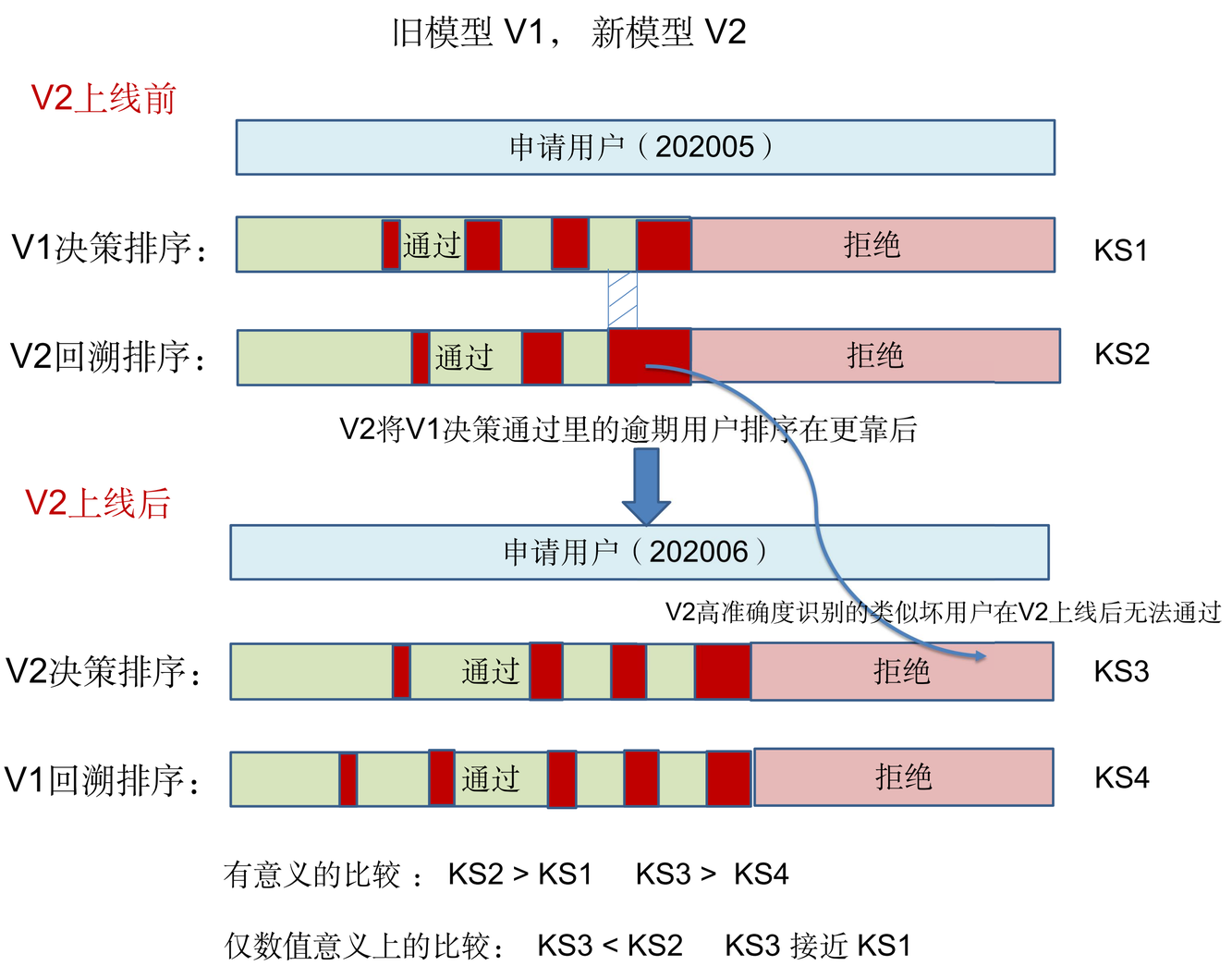
稳定性:

- 预测分数分布稳定（PSI）
- 预测分数区间对应的真实风险稳定

Swap in & out：

- 人数相同下风险下降
- 风险相同下人数提升

上线应用后，KS下降，出什么问题了?

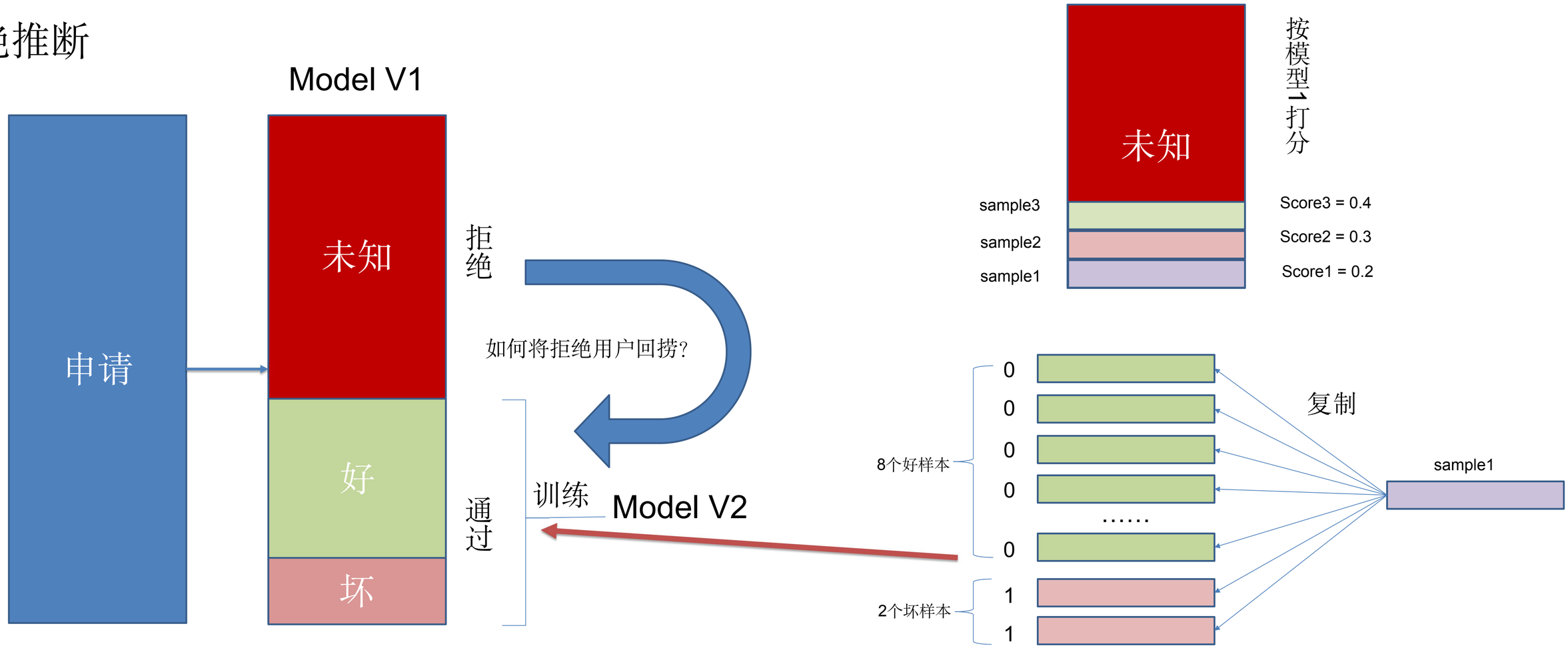


有意义的比较： KS2 > KS1 KS3 > KS4

仅数值意义上的比较： KS3 < KS2 KS3 接近 KS1

科学评估数据

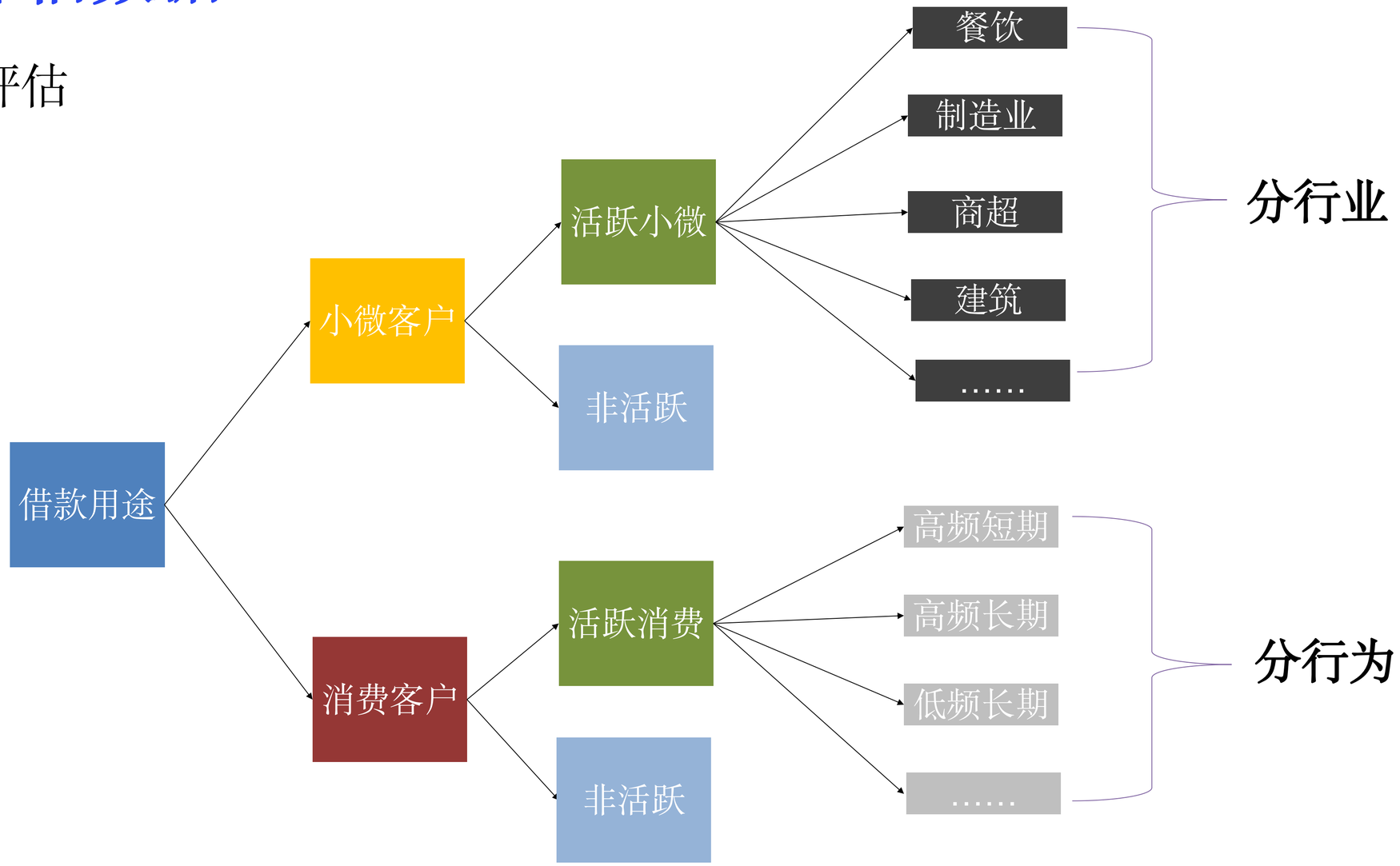
拒绝推断



增加拒绝推断，防止模型学习的样本越来越窄（增加X取值的多样性）

科学评估数据

细分客群评估



04

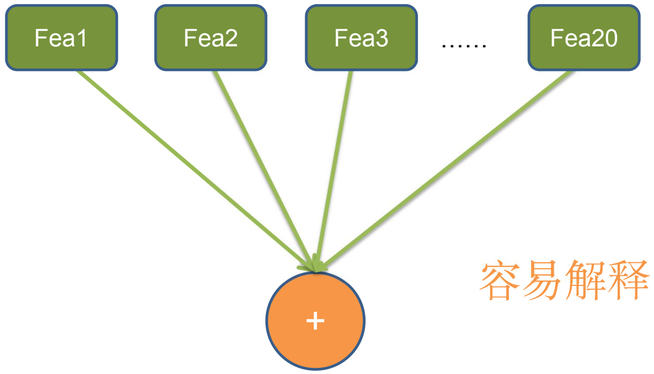
科学解释数据

Subject

如何解释模型结果？

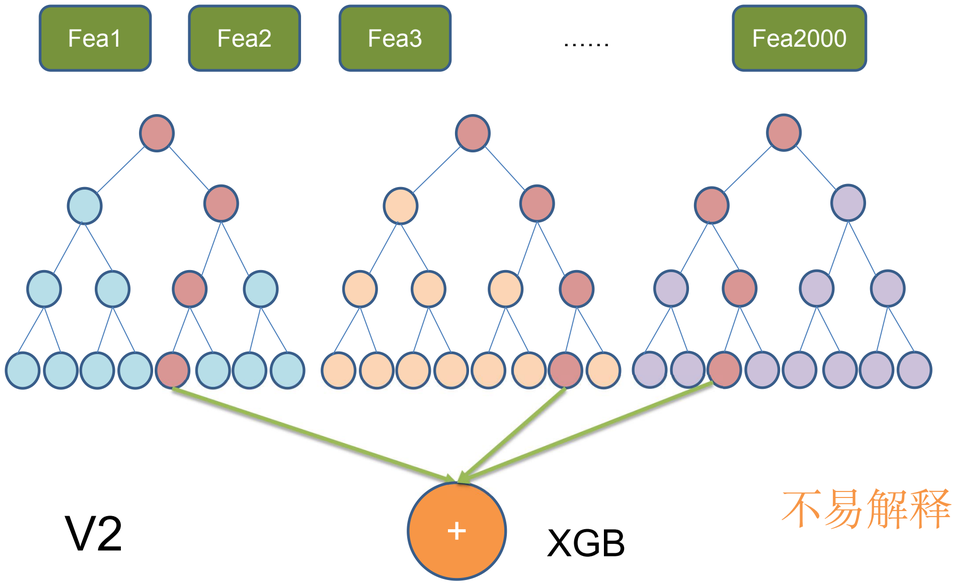
科学解释数据

V1



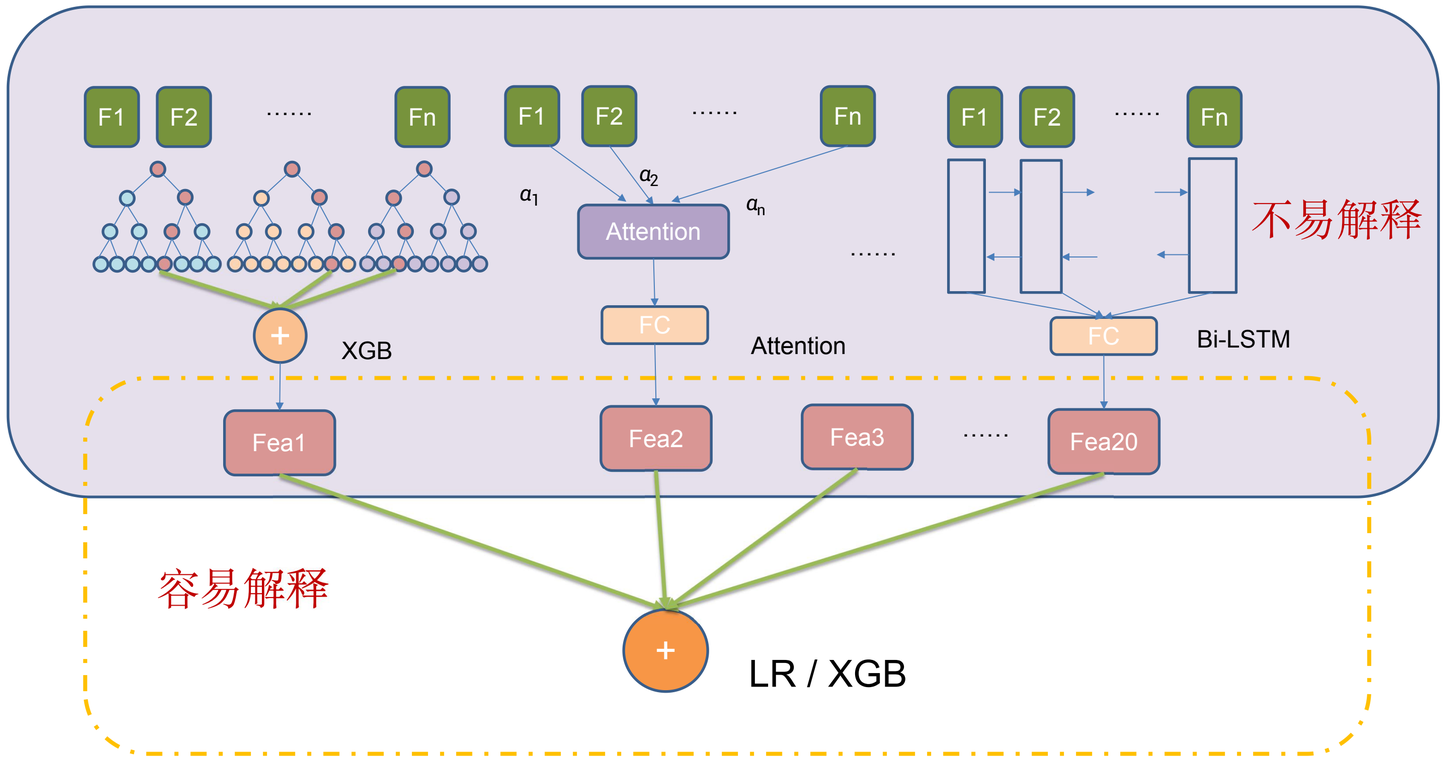
容易解释

V2



不易解释

V3



大量基础数据通过各种复杂模型(XGB,深度学习)来产生特征变量，再通过LR/浅层XGB完成最终预测模型

度小满金融 | DataFunSummit

THANKS!

今天的分享就到这里...

Ending