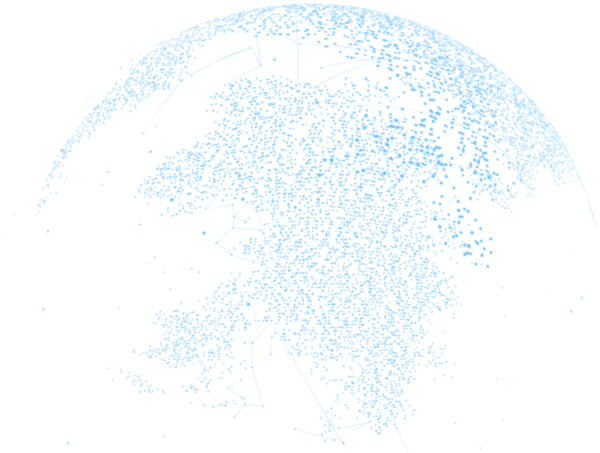
个推
GeTui

利用机器学习自动化预测用户行为



大数据服务实体产业

独角兽

领跑推送市场

个推
GeTui

最具公益力企业

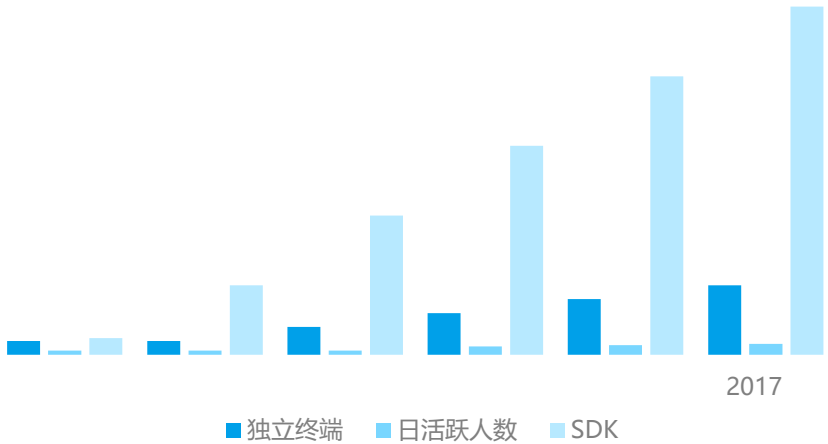
D轮融资

独立的智能大数据服务商

个推大数据规模

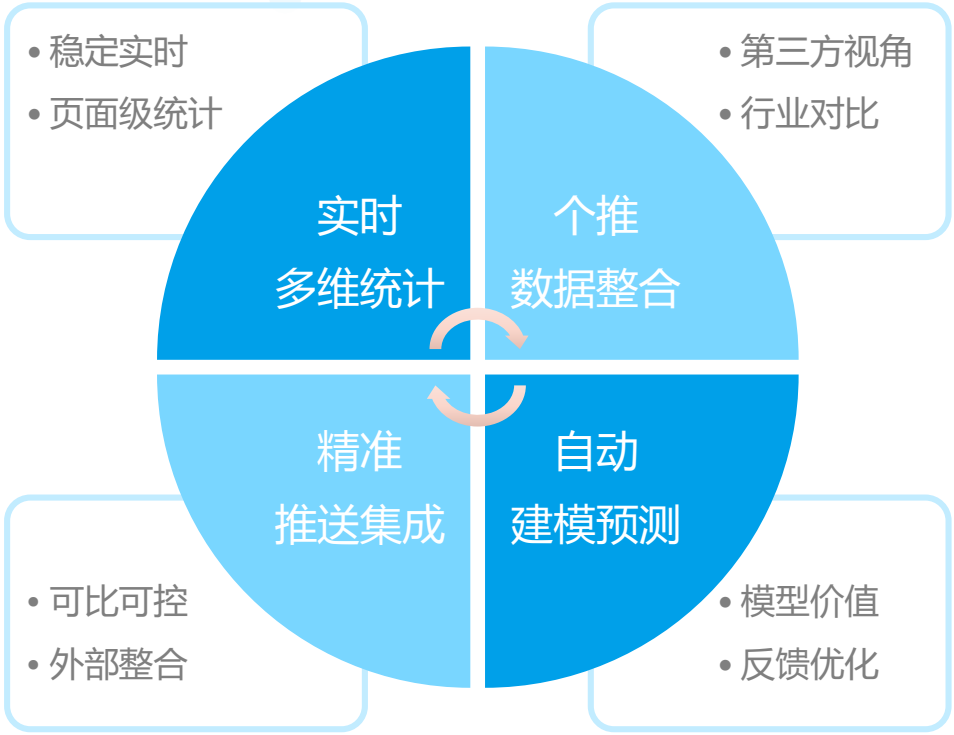
个推专注消息推送服务多年，拥有庞大的数据体系和深入的洞察能力。

个推SDK累计覆盖安装量**数百亿**，覆盖独立终端**数十亿**，服务于**数十万**APP。



个推 GeTui

个数核心能力



个推 GeTui

用户标签

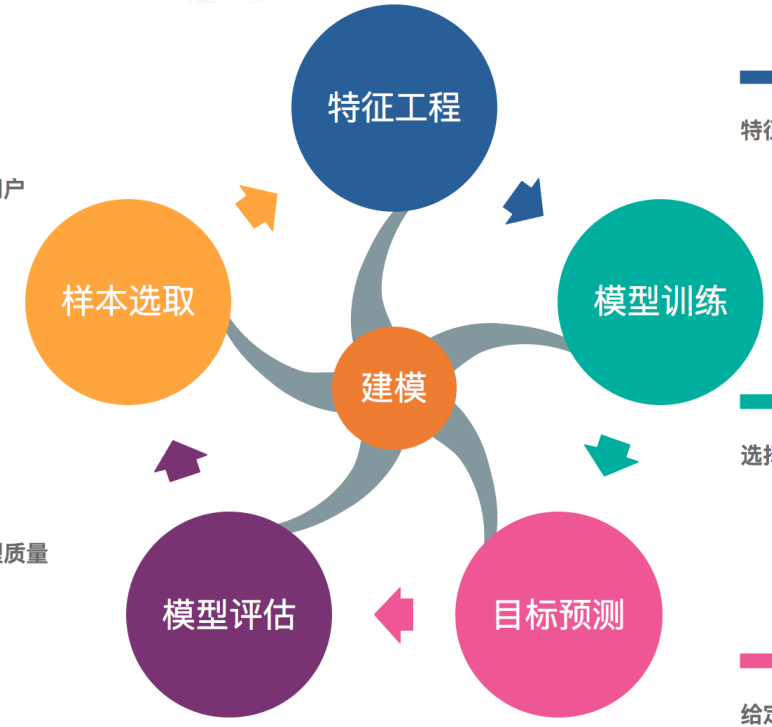
数据热度  冷：稳定  温：近期  热：实时

标签体系

基本属性	兴趣爱好	用户场景
性别	金融	旅游景区
年龄段	购物	汽车维修
婚姻	旅游	餐饮场所
有车	教育	早教中心
消费能力	汽车	口腔医院
家乡	新闻	歌剧院
职业	运动	健身中心
	办公	电器商场

个推 GeTui

流程



特征补全，合理降维

选择合适模型，配置海量计算

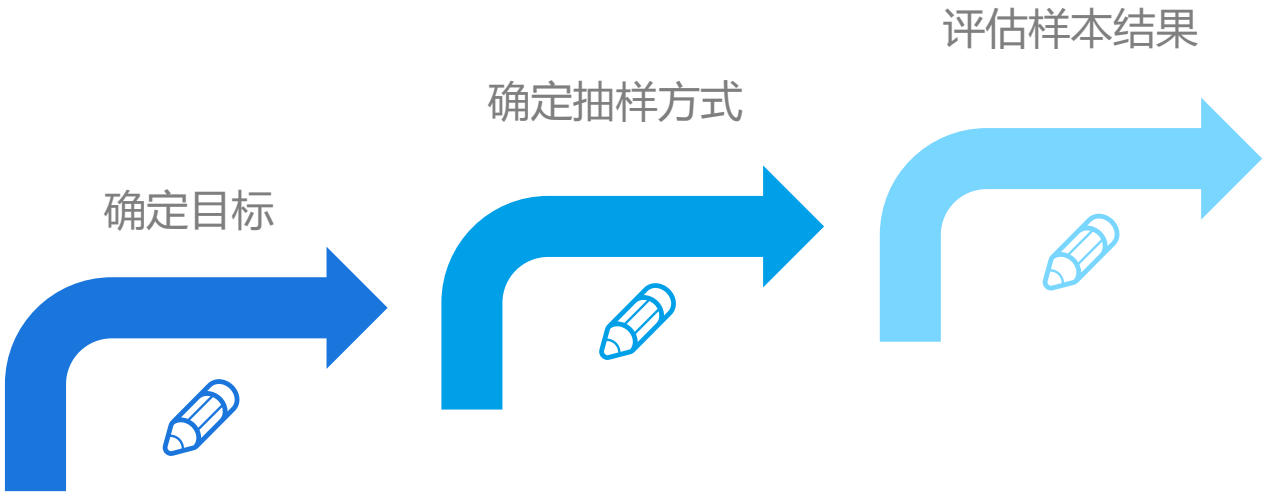
给定目标用户，预测相应的事件

多维度评估模型质量

选择正负样本用户

个推 GeTui

样本自动选取



特征自动选取

特征：安装APP， 活跃APP， 用户标签 ...

计算：特征饱和度， iv，chi，gain 等值

选择标准：综合排序以上指标，选择靠前的特征

aim	cnt_1	cnt_0	cnt_1_rate	cnt_0_rate	cnt_1/0_rate	cnt_total	cnt_rate	iv	chi	gain	iv_rank	chi_rank	gain_rank	total_rank
wtjjs42	2089	642	0.04	0.01	3.25	2731	0.06	0.074591111785068...	812.68680374	21926...|0.012759078493241...	28	16	26	70
sell_mode103	1319	249	0.03	0.01	5.3	1568	0.03	0.075986424079709...	754.69448202	19629...|0.012345732938390...	26	34	28	88
insured_age0	1886	557	0.04	0.01	3.39	2443	0.05	0.070394697368615...	761.54372400	49750...|0.011995361090997...	29	32	29	90
amount_a0	570	1883	0.01	0.04	0.3	2453	0.05	0.068167376549226...	740.44888734	48420...|0.011641396481032...	30	36	30	96
is_selfY	5742	8418	0.12	0.17	0.68	14160	0.29	0.059895756772296...	715.81651990	39660...|0.010751837006546...	31	38	31	100
commission_rate2	659	6435	0.01	0.13	0.1	7094	0.15	0.613320686033937...	5513.6177542	39974...|0.093615932200058...	1	115	1	117
agent_jd_num1.0	4410	6826	0.09	0.14	0.65	11236	0.23	0.056851066816670...	677.22007376	38900...|0.010188508400606...	33	62	33	128
agent_jd_num	4232	6502	0.09	0.13	0.65	10734	0.22	0.051815850728865...	617.42806750					

模型训练

01 开放集群

02 GPU资源队列

03 双模型架构

04 数据样本周期

05 模型算法

个推 GeTui

目标预测

01 特征补全

02 预测概率

03 TF Serving集群

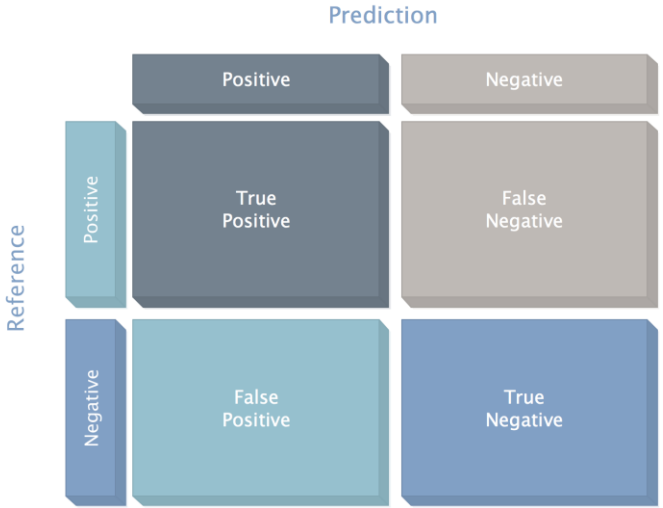
04 实时预测

05 预测失效

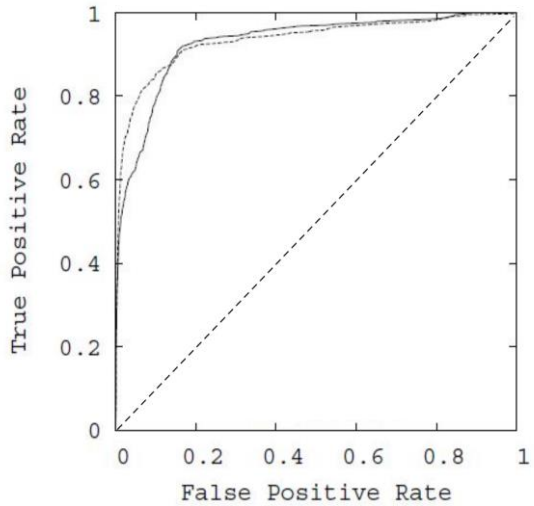
个推 GeTui

模型评估

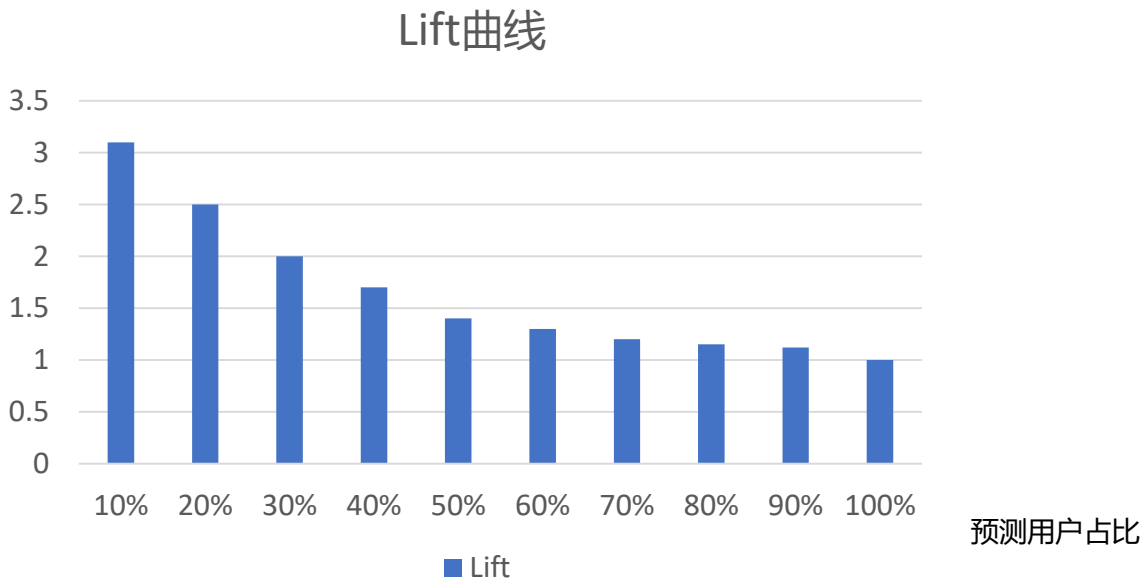
精确率、召回率



ROC、AUC



Lift曲线



模型监控

01 测试用户随机抽取，对精确率、召回率、Lift至、以及对模型AUC进行每日评估

02 预测结果全量保存1个月，用于历史回溯

03 每天重新校验预测结果，监测精确率、召回率、Lift值的变化情况

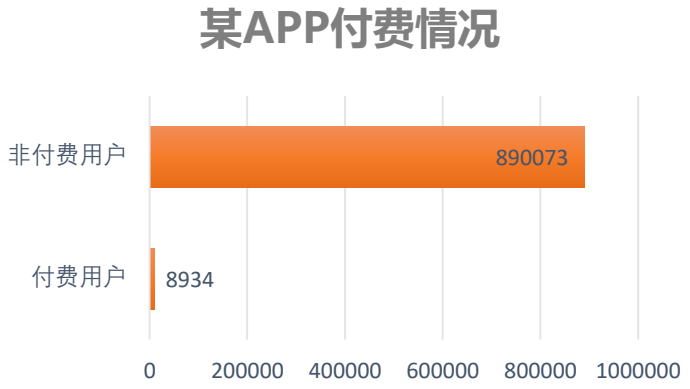
04 重要指标以报表形式呈现给开发者

个推 GeTui

挑战

样本比例
极度悬殊（付费）

重采样，分层抽样，
Bagging算法解决模型不
准问题



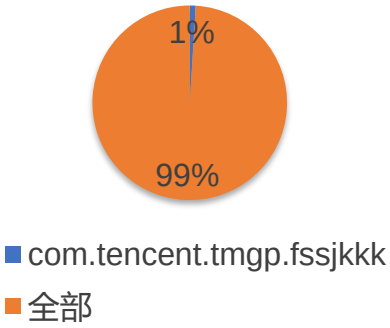
个推 GeTui

挑战

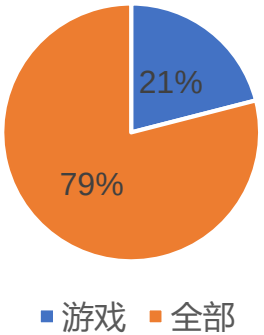
特征稀疏，饱和度低

向更高维度聚合，使特征更加饱和

包名占比



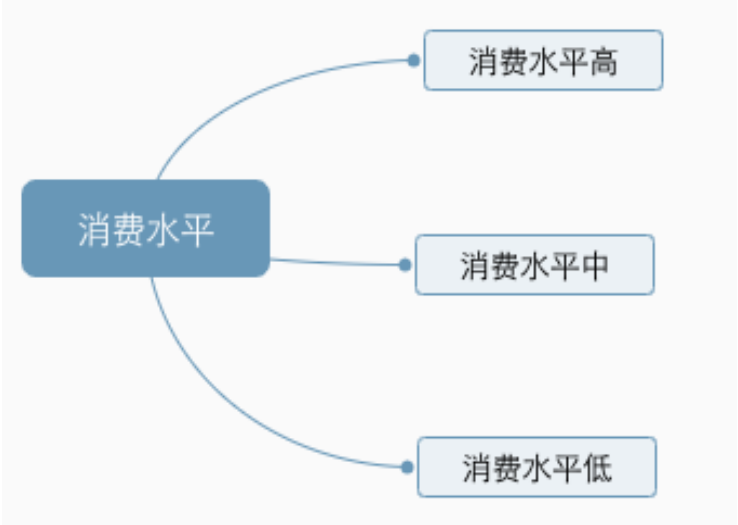
类别占比



挑战

特征中有包含关系（特征线性相关）

事先知道包含的剔除一个，不知道含义的用正则来处理一次



个推 GeTui

挑战

特征列超过2千列

用spark解决hive展开失败问题

APP每天的特征可能都会变化（发版）

每天重新调整特征和模型文件

许多百万级应用按时出结果

Az启用子线程调度集群资源训练多个模型

个推 GeTui

谢谢