训练样本集的生成

广告白名单

部门名称 | 日期

1. 训练样本集生成的主要思想
2. 回流词表的生成逻辑
3. 7天回流词表与3日回流词表的对比
4. CVR阈值的判断-过滤异常广告
5. 讨论

训练样本生成的主要思想

cvr模型

- 主要思想：
- 1、训练样本集中，有尽可能多的真正例、真负例，尽可能少的假负例。
- 2、样本量充足。

- 注：

a. 真实样本：需要预估的广告

b. 假负例：由于样本回流慢，而导致的假负例。

c. 保证样本量充足：样本增广、mock样本

回流词表的生成逻辑

以3天回流词表中一个adid为例

1、现实时间=0323

取 tomcat_ads_cv_log 中

day（servertime） between 0316 and 0322

取 clicktime = 0316 所有归因 count distinct token

2、clicktime =0316 的
1日转化数 servertime = 0316
2日转化数 servertime = 0317
3日转化数 servertime = 0318
生成 da_ad_cvtype_returnnum_3d

adid	cvtype	num1	num2	num3	day	srctype
20345103	3	240	46	18	2020-03-16	1
20345103	3	262	57	11	2020-03-17	1
20345103	3	249	24	13	2020-03-18	1
20345103	3	240	24	8	2020-03-19	1
20345103	3	216	31	6	2020-03-20	1
20345103	3	166	21	8	2020-03-21	1
20345103	3	189	33	7	2020-03-22	1

3、clicktime= 0316 ，

1日回流率（0316转化数/0316-0318转化数)

2日回流率（0316-0317转化数/0316-0318转化数)

3日回流率=1，

生成da_ad_cvtype_returnrate _3d

id	cvtype	num1	num2	num3	rate1	rate2	rate3	day	srctype	key
20345103	2	236	51	33	0.738	0.897	1.000	2020-03-16	1	adid
20345103	2	270	74	25	0.732	0.932	1.000	2020-03-17	1	adid
20345103	2	261	63	14	0.772	0.959	1.000	2020-03-18	1	adid
20345103	2	245	61	17	0.759	0.947	1.000	2020-03-19	1	adid
20345103	2	235	73	15	0.728	0.954	1.000	2020-03-20	1	adid
20345103	2	195	53	19	0.730	0.929	1.000	2020-03-21	1	adid
20345103	2	190	46	14	0.760	0.944	1.000	2020-03-22	1	adid

4、clicktime =0322（此条出现的时间是clicktime）

1日加权平均回流率（0316-0322 的 1日回流率 * 0316-0322的 1日转化数权重）

2日加权平均回流率（ 0316-0322 的 2日回流率 * 0316-0322 的 （1日转化数+2日转化数）权重）

1日变异系数= 总体标准差（ 0316-0322 的 1日回流率）/ 平均值（ 0316-03220的1日回流率）

2日变异系数= 总体标准差（ 0316-0322 的 2日回流率）/ 平均值（ 0316-0322）的2日回流率）

id	cvtype	avg_rate1	avg_rate2	avg_rate3	cof_variation1	cof_variation2	cof_variation3	ndays	covnum	maxcovday	day	srctype	key
20345103	2	0.75	0.93	1.00	0.03342	0.02168	0	7	1078	2020-03-16	2020-03-16	1	adid
20345103	2	0.75	0.93	1.00	0.03514	0.02162	0	7	1323	2020-03-17	2020-03-17	1	adid
20345103	2	0.75	0.93	1.00	0.03234	0.02385	0	7	1535	2020-03-18	2020-03-18	1	adid
20345103	2	0.75	0.93	1.00	0.02958	0.02378	0	7	1723	2020-03-19	2020-03-19	1	adid
20345103	2	0.75	0.94	1.00	0.02641	0.02273	0	7	1920	2020-03-20	2020-03-20	1	adid
20345103	2	0.75	0.94	1.00	0.02462	0.02037	0	7	2054	2020-03-21	2020-03-21	1	adid
20345103	2	0.75	0.94	1.00	0.02209	0.02054	0	7	2190	2020-03-22	2020-03-22	1	adid

三日回流词表与七日回流词表对比

三日回流词表比七日回流词表快

3日回流

	1日回流率分组的广告计数占比		2日回流率分组的广告计数占比	
	自有流量	联盟流量	自有流量	联盟流量
下载	100.0%	100.0%	100.0%	100.0%
<0.8	0.8%	0.6%	0.2%	0.1%
0.8-0.9	0.7%	0.8%	0.1%	0.1%
0.9-1	15.7%	34.7%	9.2%	20.7%
=1	82.8%	63.9%	90.5%	79.1%
注册	100.0%	100.0%	100.0%	100.0%
<0.8	47.1%	65.8%	8.2%	10.2%
0.8-0.9	19.0%	17.9%	14.3%	30.3%
0.9-1	4.5%	4.2%	20.3%	31.4%
=1	29.5%	12.1%	57.2%	28.1%
首次激活	100.0%	100.0%	100.0%	100.0%
<0.8	59.2%	54.7%	10.1%	4.8%
0.8-0.9	11.3%	24.0%	17.0%	16.5%
0.9-1	2.8%	5.2%	20.3%	44.2%
=1	26.7%	16.1%	52.6%	34.5%
表单	100.0%		100.0%	
<0.8	0.3%			
0.8-0.9	0.4%			
0.9-1	5.3%		0.8%	
=1	94.0%		99.2%	
自定义激活	100.0%	100.0%	100.0%	100.0%
<0.8	23.1%	32.4%	0.8%	1.0%
0.8-0.9	16.4%	17.3%	1.3%	3.8%
0.9-1	17.1%	24.5%	9.0%	24.6%
=1	43.4%	25.8%	88.9%	70.6%
自定义注册	100.0%	100.0%	100.0%	100.0%
<0.8	17.4%	59.9%	1.6%	0.0%
0.8-0.9	11.6%	3.8%	1.9%	1.4%
0.9-1	16.0%	9.9%	13.5%	2.8%
=1	55.1%	26.4%	83.1%	95.8%

7日回流

	1日回流率分组的广告计数占比		3日回流率分组的广告计数占比	
	自有流量	联盟流量	自有流量	联盟流量
下载	100.0%	100.0%	100.0%	100.0%
<0.8	1.7%	1.0%	0.3%	0.2%
0.8-0.9	2.4%	1.1%	0.4%	0.2%
0.9-1	27.7%	40.0%	16.0%	21.8%
1	68.2%	58.0%	83.4%	77.9%
注册	100.0%	100.0%	100.0%	100.0%
<0.8	56.9%	75.8%	8.1%	12.2%
0.8-0.9	14.5%	12.9%	15.2%	34.1%
0.9-1	2.6%	1.9%	20.9%	29.5%
1	26.1%	9.4%	55.8%	24.2%
首次激活	100.0%	100.0%	100.0%	100.0%
<0.8	70.9%	71.5%	20.5%	9.0%
0.8-0.9	7.4%	13.8%	22.6%	31.2%
0.9-1	1.8%	2.2%	13.3%	32.7%
1	19.8%	12.5%	43.6%	27.1%
表单	100.0%		100.0%	
<0.8	0.6%		0.001067236	
0.8-0.9	0.7%		0.001067236	
0.9-1	9.7%		2.1%	
1	88.9%		97.7%	
自定义激活	100.0%	100.0%	100.0%	100.0%
<0.8	25.6%	35.8%	1.4%	1.8%
0.8-0.9	17.2%	18.7%	2.1%	4.9%
0.9-1	16.9%	23.8%	12.6%	29.6%
1	40.3%	21.6%	83.8%	63.7%
自定义注册	100.0%	100.0%	100.0%	100.0%
<0.8	19.9%	60.3%	1.8%	0.0%
0.8-0.9	14.9%	1.3%	2.0%	0.0%
0.9-1	11.0%	14.6%	14.7%	7.3%
1	54.2%	23.8%	81.5%	92.7%

训练样本集覆盖率

三日回流词表与7日回流词表对比

名词解释：

1、**用于模型训练的样本集 （以下简称 训练样本集）**：通过方法生成的样本集，是人为干扰构建的。

例：

① 取0323三日回流词表+过滤逻辑后生成的adid-list，Join 所有0325点击数、0325广告数，0325-0327归因 if(cvtype in (2,3)),1日加权平均回流率>0.8, 1日加权平均回流率>0.9)

② 取0323三日回流词表+过滤逻辑后生成的adid-list，Join 所有0323点击数、0323广告数，0323-0325归因 if(cvtype in (2,3)),3日加权平均回流率>0.8, 3日加权平均回流率>0.9)

③ 取0319七日回流词表+过滤逻辑后生成的adid-list，Join 所有0323点击数、0323广告数，0323-0325归因 if(cvtype in (2,3)),3日加权平均回流率>0.8, 3日加权平均回流率>0.9)

④ 取0319七日回流词表+过滤逻辑后生成的adid-list，Join 所有0325点击数、0325广告数，0325-0327归因 if(cvtype in (2,3)),1日加权平均回流率>0.8, 1日加权平均回流率>0.9)

2、**Mock样本集**：adid+reqid粒度上，样本的转化目标与adid当前的转化目标不同的集合

3、**线上真实的预估过的oCPC广告样本集（以下简称 真实样本集）**： cvr模型需要预估的、对应优化目标的那部分oCPC广告，产生的后验转化数、广告数 、点击数

例：

① 取0325日，ocpc_stage >1时，不同阶段对应的广告优化目标的对应的转化数、广告数、点击数

② 取0323日，ocpc_stage >1时，不同阶段对应的广告优化目标的对应的转化数、广告数、点击数

4、有归因的广告list对应的集合（以下简称 全量 归因样本集）：

例：

① 取0325日点击对应的，0325-0327的归因数、0325日所有adid的点击数，广告数

② 取0323日点击对应的，0323-0325的归因数、0323日所有adid的点击数，广告数

5、训练样本集join真实样本集：（交集）

取数口径的问题，可以忽略不计。构建的adid-list中取ocpc_conversion_type，除deep_type=2的情况，作为广告真实集的标签。

由于没有使用ocpc_stage，这里的真实样本集和上文中有区别。在游戏注册上差的多一些。

训练样本集覆盖率-浏览器推荐页

3日回流词表中真实样本比例高，3日回流词表较好

3天

1、训练样本集/归因的集合

0325	正例数	点击数	广告数
下载	82%	71%	55%
游戏注册	20%	15%	13%
首次激活	12%	7%	7%
表单提交	67%	42%	47%
自定义激活	44%	30%	24%
自定义注册	72%	53%	34%

2、训练样本集join 真实样本集/训练样本集

0325	正例数	点击数	广告数
下载	26%	25%	28%
游戏注册	93%	91%	94%
首次激活	57%	33%	14%
表单提交	99%	96%	99%
自定义激活	61%	57%	60%
自定义注册	54%	53%	49%

3、训练样本集join 真实样本集/真实样本集

0325	正例数	点击数	广告数
下载	71%	67%	28%
游戏注册	21%	20%	13%
首次激活	11%	8%	3%
表单提交	70%	54%	54%
自定义激活	56%	38%	27%
自定义注册	81%	66%	20%

7天

1、训练样本集/归因的集合

0325	正例数	点击数	广告数
下载	72%	52%	45%
游戏注册	7%	4%	5%
首次激活	2%	2%	3%
表单提交	45%	30%	40%
自定义激活	33%	23%	17%
自定义注册	48%	34%	20%

2、训练样本集join 真实样本集/训练样本集

0325	正例数	点击数	广告数
下载	31%	39%	58%
游戏注册	95%	89%	93%
首次激活	0%	0%	0%
表单提交	99%	94%	99%
自定义激活	73%	61%	63%
自定义注册	70%	70%	66%

3、训练样本集join 真实样本集/真实样本集

0325	正例数	点击数	广告数
下载	75%	77%	48%
游戏注册	7%	6%	5%
首次激活	0%	0%	0%
表单提交	47%	39%	46%
自定义激活	50%	32%	20%
自定义注册	68%	56%	16%

构建样本集覆盖率-浏览器推荐页

3日回流词表中真实样本比例高，3日回流词表较好

3天

1、训练样本集/归因的集合

0323	正例数	点击数	广告数
下载	86%	73%	63%
游戏注册	69%	52%	28%
首次激活	73%	49%	25%
表单提交	88%	68%	65%
自定义激活	75%	65%	39%
自定义注册	85%	70%	48%

2、训练样本集join 真实样本集/训练样本集

0323	正例数	点击数	广告数
下载	25%	20%	27%
游戏注册	90%	88%	89%
首次激活	73%	43%	35%
表单提交	99%	96%	99%
自定义激活	57%	51%	56%
自定义注册	46%	55%	45%

3、训练样本集join 真实样本集/真实样本集

0323	正例数	点击数	广告数
下载	72%	69%	34%
游戏注册	69%	64%	27%
首次激活	79%	74%	35%
表单提交	95%	89%	66%
自定义激活	86%	75%	37%
自定义注册	89%	77%	26%

7天

1、训练样本集/归因的集合

0323	正例数	点击数	广告数
下载	75%	46%	49%
游戏注册	48%	39%	18%
首次激活	63%	36%	19%
表单提交	58%	44%	52%
自定义激活	61%	49%	29%
自定义注册	70%	56%	34%

2、训练样本集join 真实样本集/训练样本集

0323	正例数	点击数	广告数
下载	26%	28%	30%
游戏注册	96%	94%	94%
首次激活	77%	53%	37%
表单提交	98%	94%	99%
自定义激活	55%	45%	51%
自定义注册	46%	51%	40%

3、训练样本集join 真实样本集/真实样本集

0323	正例数	点击数	广告数
下载	67%	62%	30%
游戏注册	52%	51%	19%
首次激活	72%	67%	28%
表单提交	62%	56%	53%
自定义激活	68%	50%	25%
自定义注册	73%	57%	17%

训练样本集覆盖率-广告联盟

3日回流词表中真实样本比例高，3日回流词表较好

3天

1、训练样本集/归因的集合

0325	正例数	点击数	广告数
下载	85%	72%	58%
游戏注册	14%	7%	8%
首次激活	18%	29%	16%
自定义激活	21%	36%	15%
自定义注册	19%	5%	7%

2、训练样本集join 真实样本集/训练样本集

0325	正例数	点击数	广告数
下载	38%	31%	29%
游戏注册	90%	85%	92%
首次激活	20%	4%	23%
自定义激活	26%	5%	55%
自定义注册	6%	6%	40%

3、训练样本集join 真实样本集/真实样本集

0325	正例数	点击数	广告数
下载	87%	87%	48%
游戏注册	14%	6%	10%
首次激活	10%	10%	19%
自定义激活	13%	6%	18%
自定义注册	12%	14%	33%

7天

1、训练样本集/归因的集合

0325	正例数	点击数	广告数
下载	70%	59%	47%
游戏注册	3%	2%	2%
首次激活	4%	5%	5%
自定义激活	13%	33%	11%
自定义注册	18%	5%	4%

2、训练样本集join 真实样本集/训练样本集

0325	正例数	点击数	广告数
下载	42%	33%	32%
游戏注册	62%	54%	79%
首次激活	20%	4%	14%
自定义激活	17%	2%	51%
自定义注册	0%	0%	0%

3、训练样本集join 真实样本集/真实样本集

0325	正例数	点击数	广告数
下载	78%	77%	42%
游戏注册	2%	1%	2%
首次激活	2%	1%	4%
自定义激活	5%	2%	12%
自定义注册	0%	0%	0%

训练样本集覆盖率-广告联盟

3日回流词表中真实样本比例高，3日回流词表较好

3天

1、训练样本集/归因的集合

0323	正例数	点击数	广告数
下载	96%	84%	71%
游戏注册	77%	78%	42%
首次激活	82%	70%	41%
自定义激活	74%	75%	35%
自定义注册	71%	70%	37%

2、训练样本集join 真实样本集/训练样本集

0323	正例数	点击数	广告数
下载	42%	32%	28%
游戏注册	96%	96%	96%
首次激活	37%	15%	27%
自定义激活	45%	30%	65%
自定义注册	10%	2%	27%

3、训练样本集join 真实样本集/真实样本集

0323	正例数	点击数	广告数
下载	97%	97%	58%
游戏注册	79%	82%	47%
首次激活	85%	86%	58%
自定义激活	81%	78%	50%
自定义注册	52%	48%	50%

7天

1、训练样本集/归因的集合

0323	正例数	点击数	广告数
下载	78%	69%	54%
游戏注册	53%	59%	29%
首次激活	66%	59%	30%
自定义激活	45%	59%	24%
自定义注册	45%	41%	17%

2、训练样本集join 真实样本集/训练样本集

0323	正例数	点击数	广告数
下载	43%	33%	30%
游戏注册	94%	96%	94%
首次激活	38%	16%	29%
自定义激活	42%	23%	61%
自定义注册	0%	0%	0%

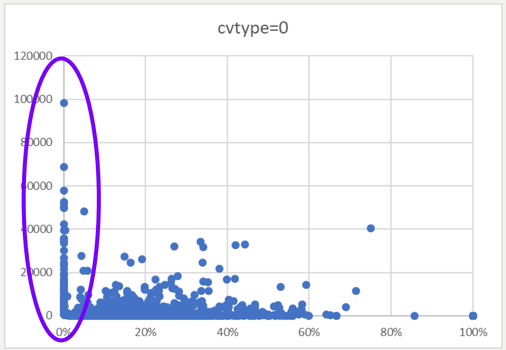
3、训练样本集join 真实样本集/真实样本集

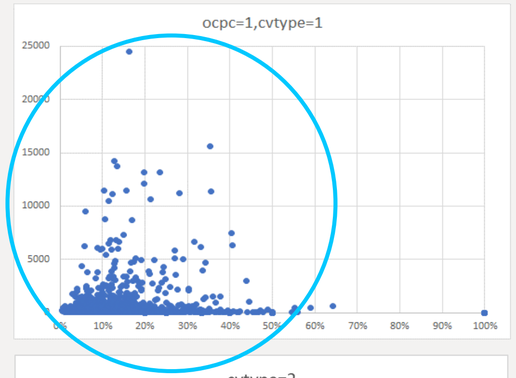
0323	正例数	点击数	广告数
下载	81%	81%	48%
游戏注册	53%	61%	32%
首次激活	70%	74%	47%
自定义激活	46%	47%	33%
自定义注册	0%	0%	0%

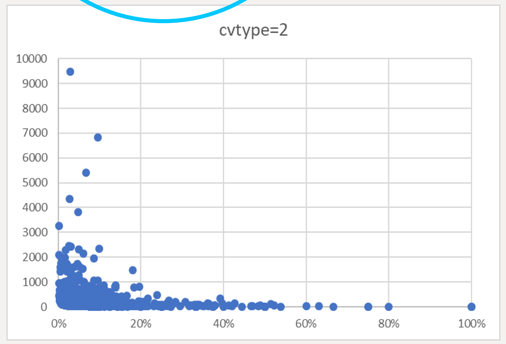
异常adid的判断

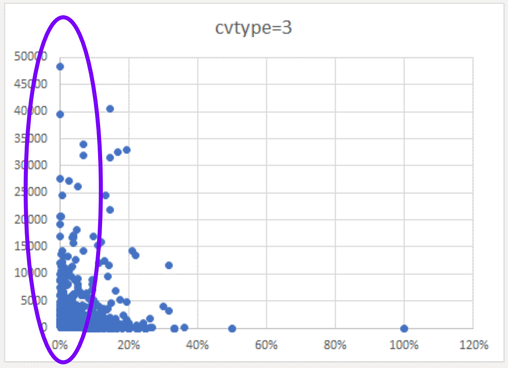
高点击低转化的adid

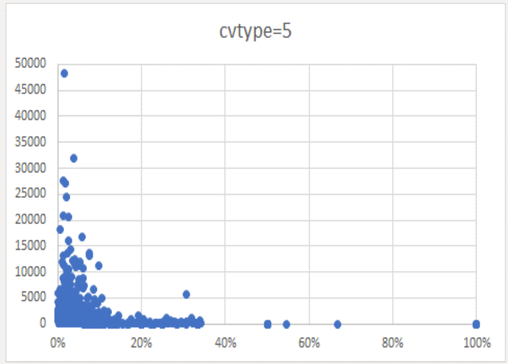
浏览器推荐页



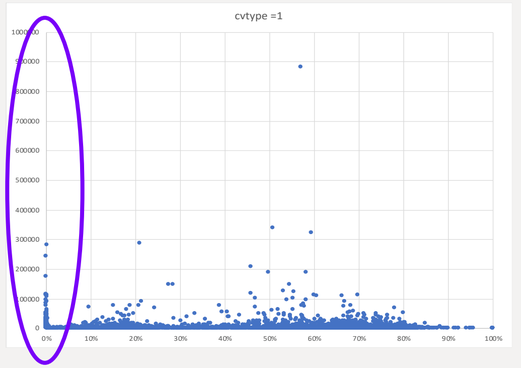


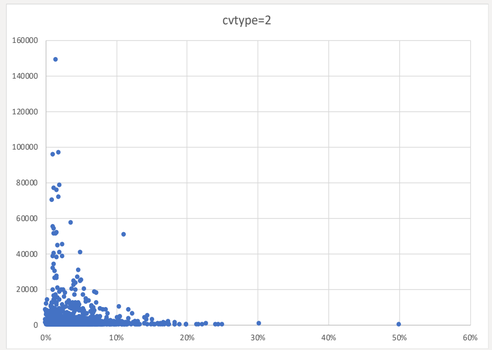


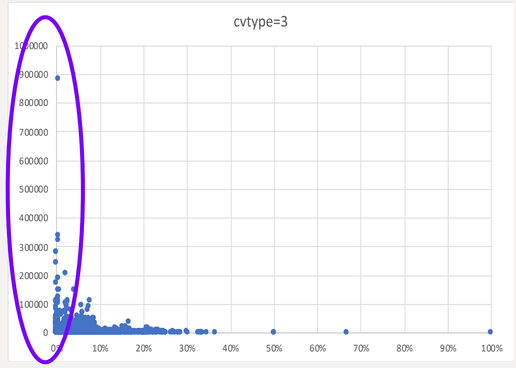


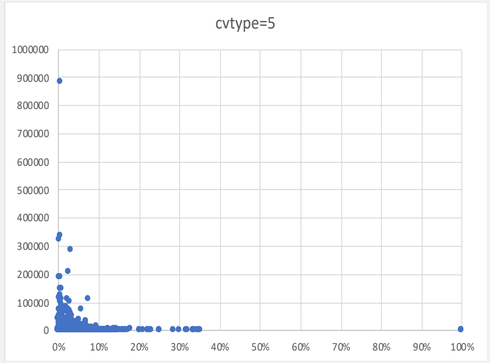


广告联盟



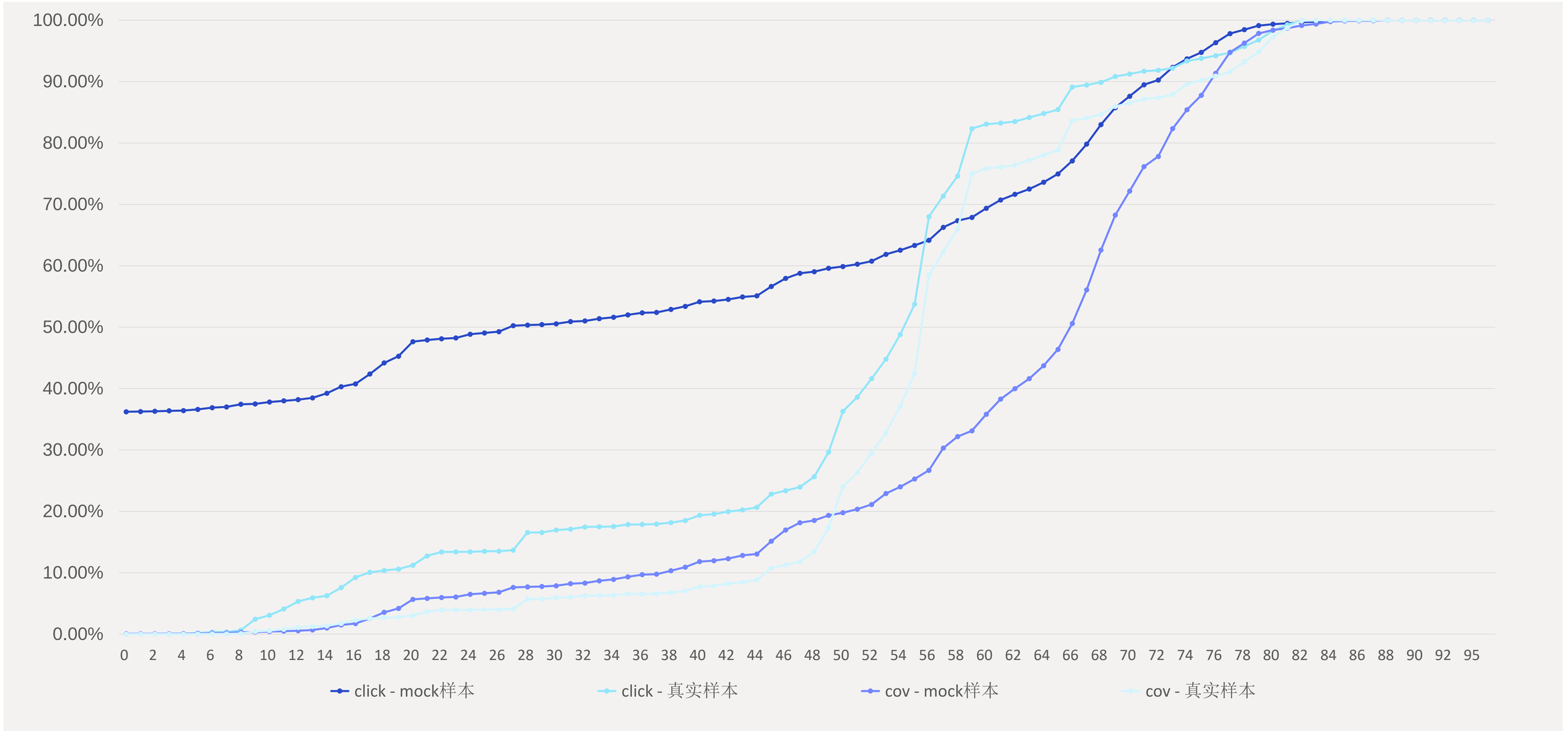






广告联盟

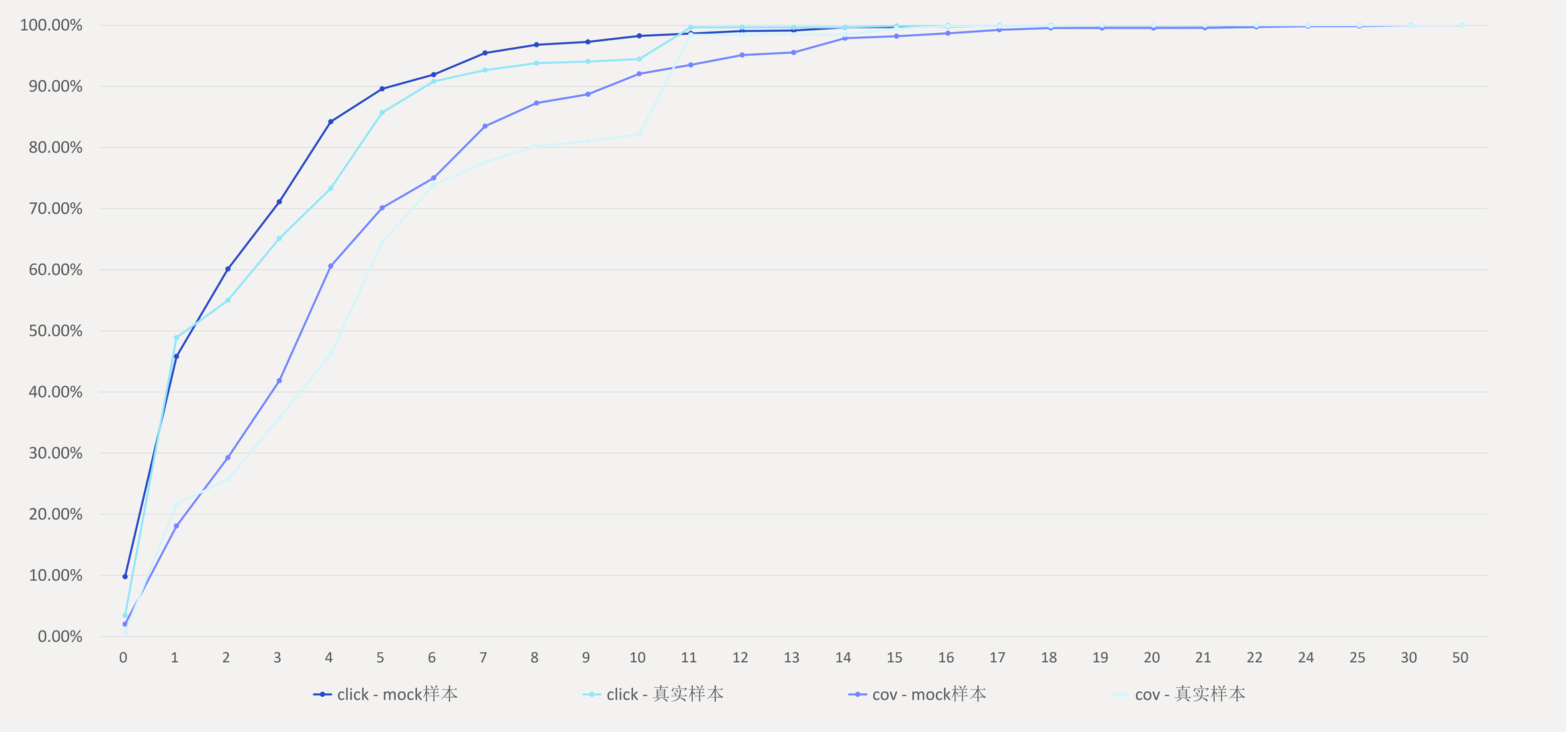
下载



横坐标：cvr分桶，0代表小于1%，1代表小于2%....
纵坐标：累计值/汇总值

广告联盟

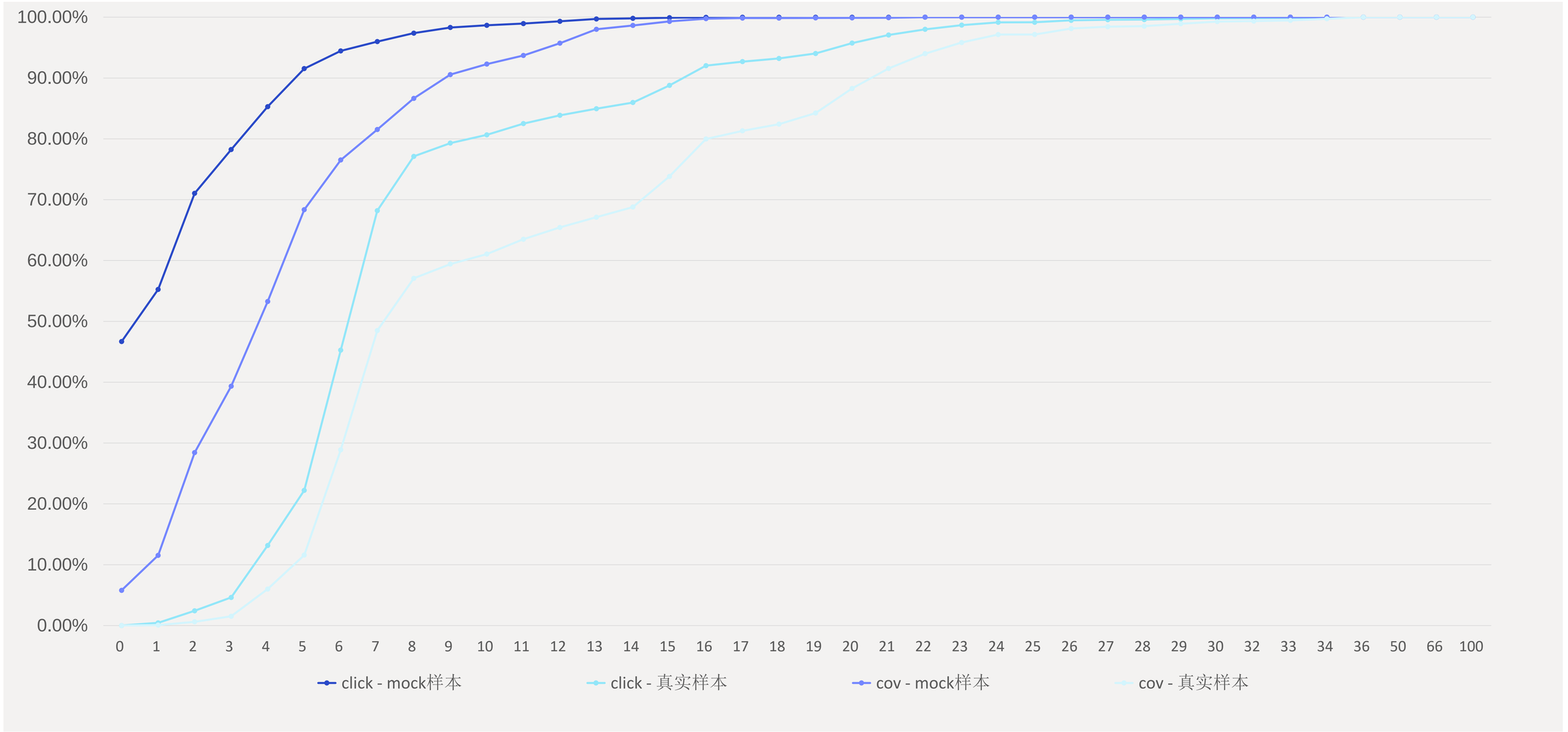
游戏注册



横坐标：cvr分桶，0代表小于1%，1代表小于2%....
纵坐标：累计值/汇总值

广告联盟

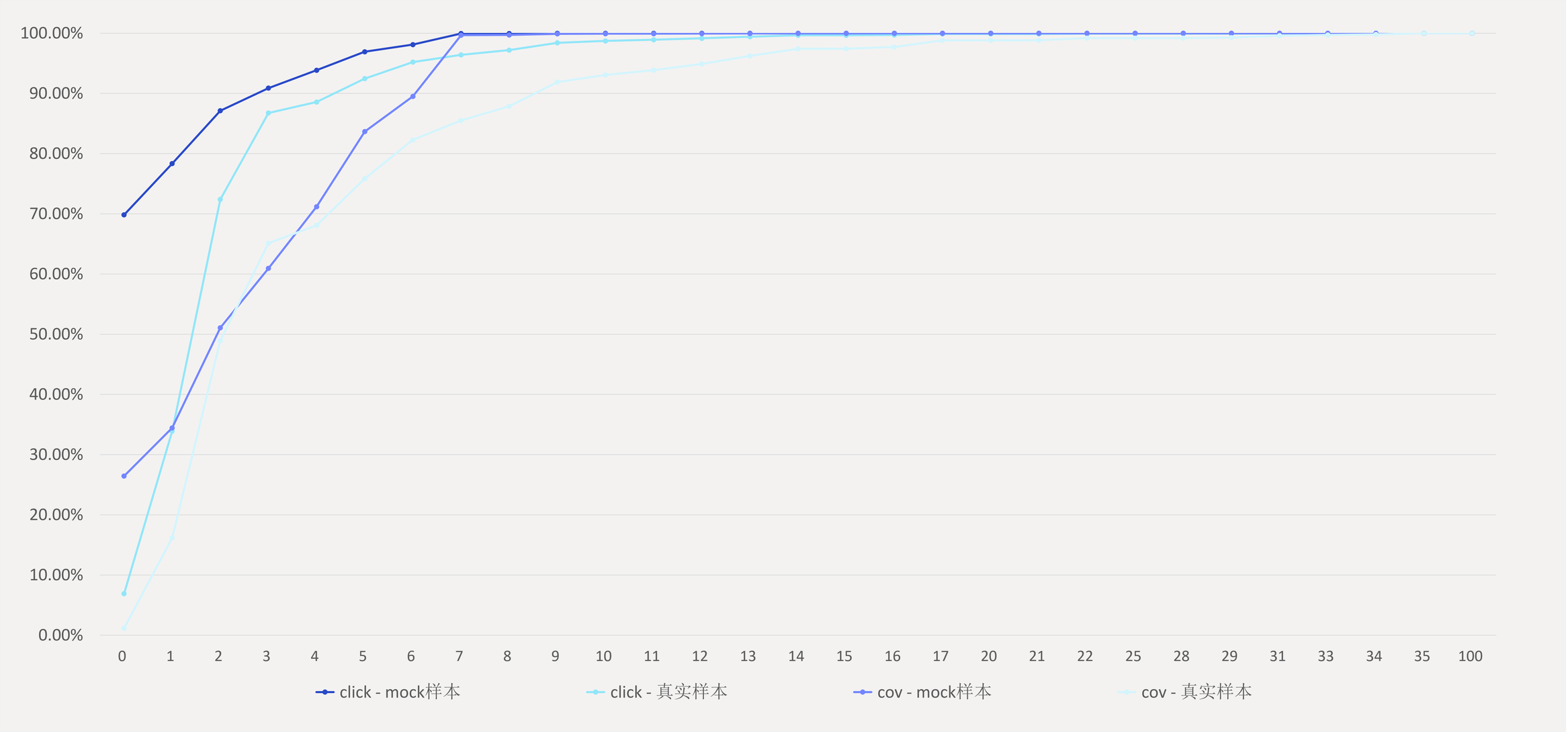
首次激活



横坐标：cvr分桶，0代表小于1%，1代表小于2%....
纵坐标：累计值/汇总值

广告联盟

自定义激活



横坐标：cvr分桶，0代表小于1%，1代表小于2%....
纵坐标：累计值/汇总值

广告联盟阈值

使用最近一个月的每周六的数据

■ **阈值设置的条件**：最近一个月，通过配置不同的阈值，使得归因结果中过滤ocpc点击/总ocpc点击，最大不超过1%，最好在0.5%以下

- **参考1**

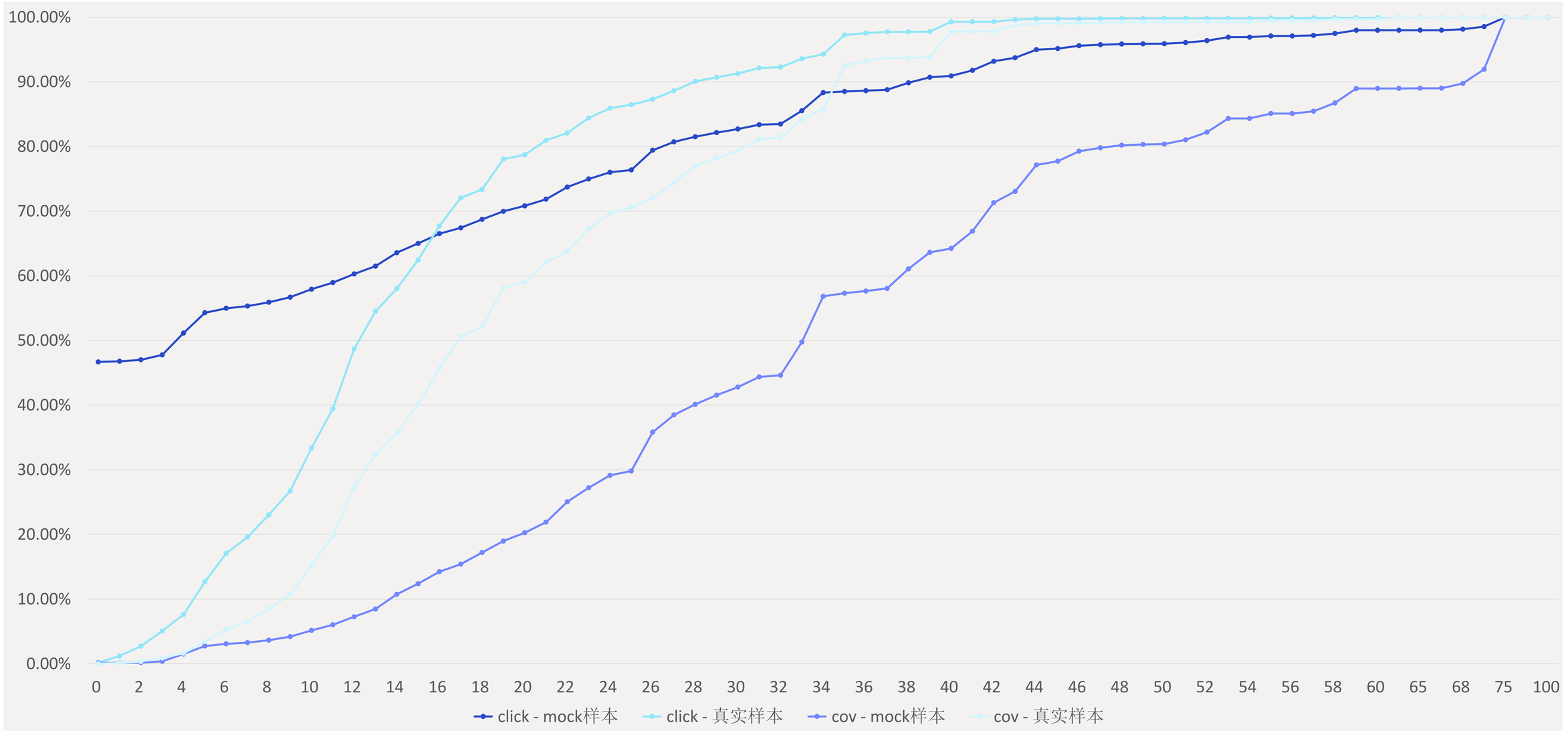
		归因结果中过滤ocpc点击/总ocpc点击				归因结果中总过滤点击/总点击			
cvr阈值	**广告联盟**	**2020/2/29**	**2020/3/7**	**2020/3/14**	**2020/3/21**	**2020/2/29**	**2020/3/7**	**2020/3/14**	**2020/3/21**
3.00%	下载	0.00%	0.01%	0.01%	0.01%	19%	27%	25%	27%
0.30%	游戏注册	0.17%	0.09%	0.21%	0.22%	3%	2%	3%	2%
1.50%	首次激活	0.31%	0.12%	1.32%	0.28%	47%	49%	42%	42%
0.16%	自定义激活	0.43%	0.00%	0.71%	0.53%	0%	0%	1%	1%

- **参考2**

		归因结果中过滤ocpc点击/总ocpc点击				归因结果中总过滤点击/总点击			
cvr阈值	**广告联盟**	**2020/2/29**	**2020/3/7**	**2020/3/14**	**2020/3/21**	**2020/2/29**	**2020/3/7**	**2020/3/14**	**2020/3/21**
7.00%	**下载**	**0.30%**	**0.18%**	**0.30%**	**0.24%**	**20%**	**27%**	**26%**	**27%**
0.50%	游戏注册	0.37%	0.09%	0.33%	0.77%	3%	3%	4%	3%
1.20%	**首次激活**	**0.29%**	**0.01%**	**1.00%**	**0.05%**	**44%**	**46%**	**40%**	**41%**
0.18%	自定义激活	0.47%	0.00%	0.71%	1.01%	1%	0%	1%	1%
0.16%	自定义注册	其oCPC广告转化率最高				67%	89%	71%	87%

浏览器推荐页

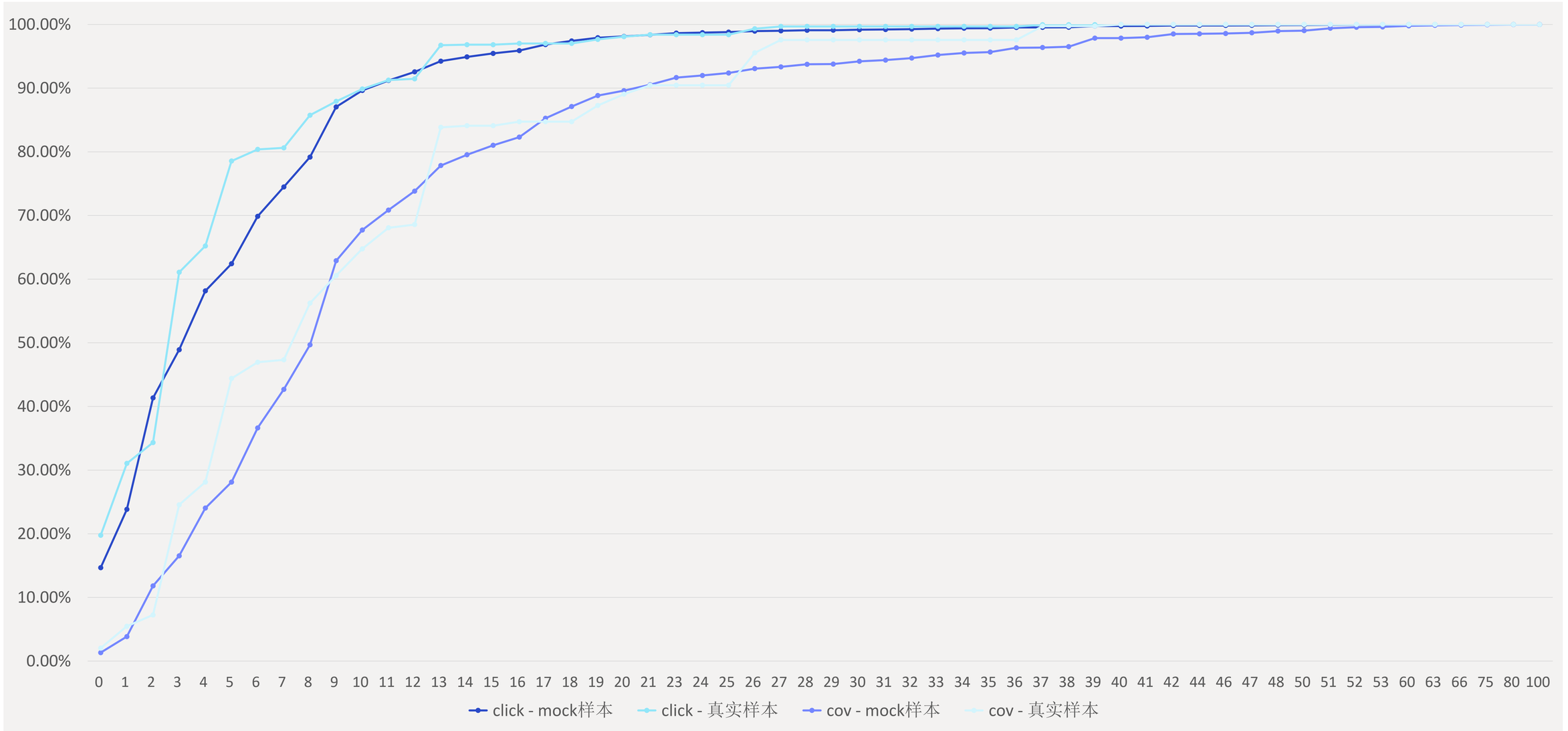
下载



横坐标：cvr分桶，0代表小于1%，1代表小于2%....
纵坐标：累计值/汇总值

浏览器推荐页

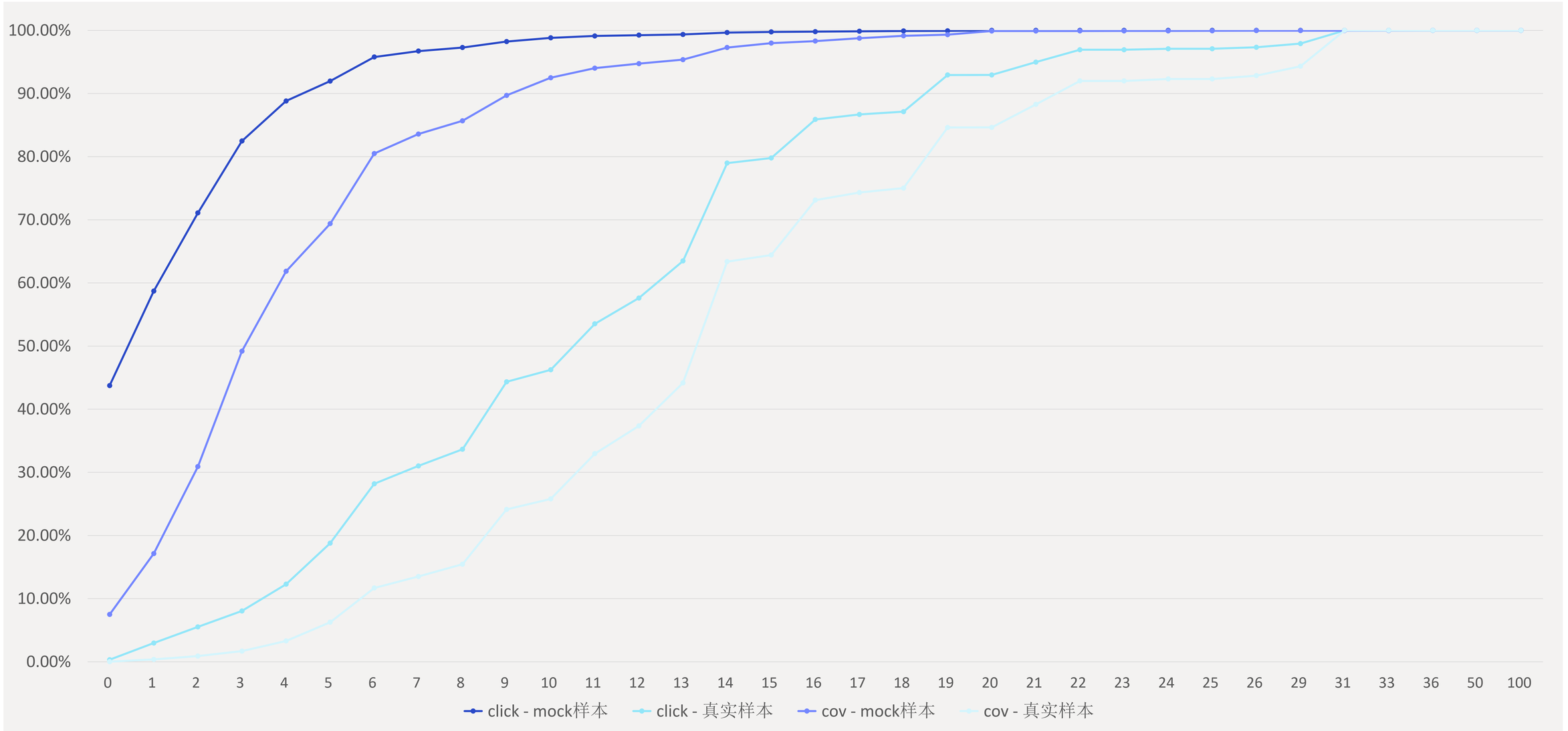
游戏注册



横坐标：cvr分桶，0代表小于1%，1代表小于2%....
纵坐标：累计值/汇总值

浏览器推荐页

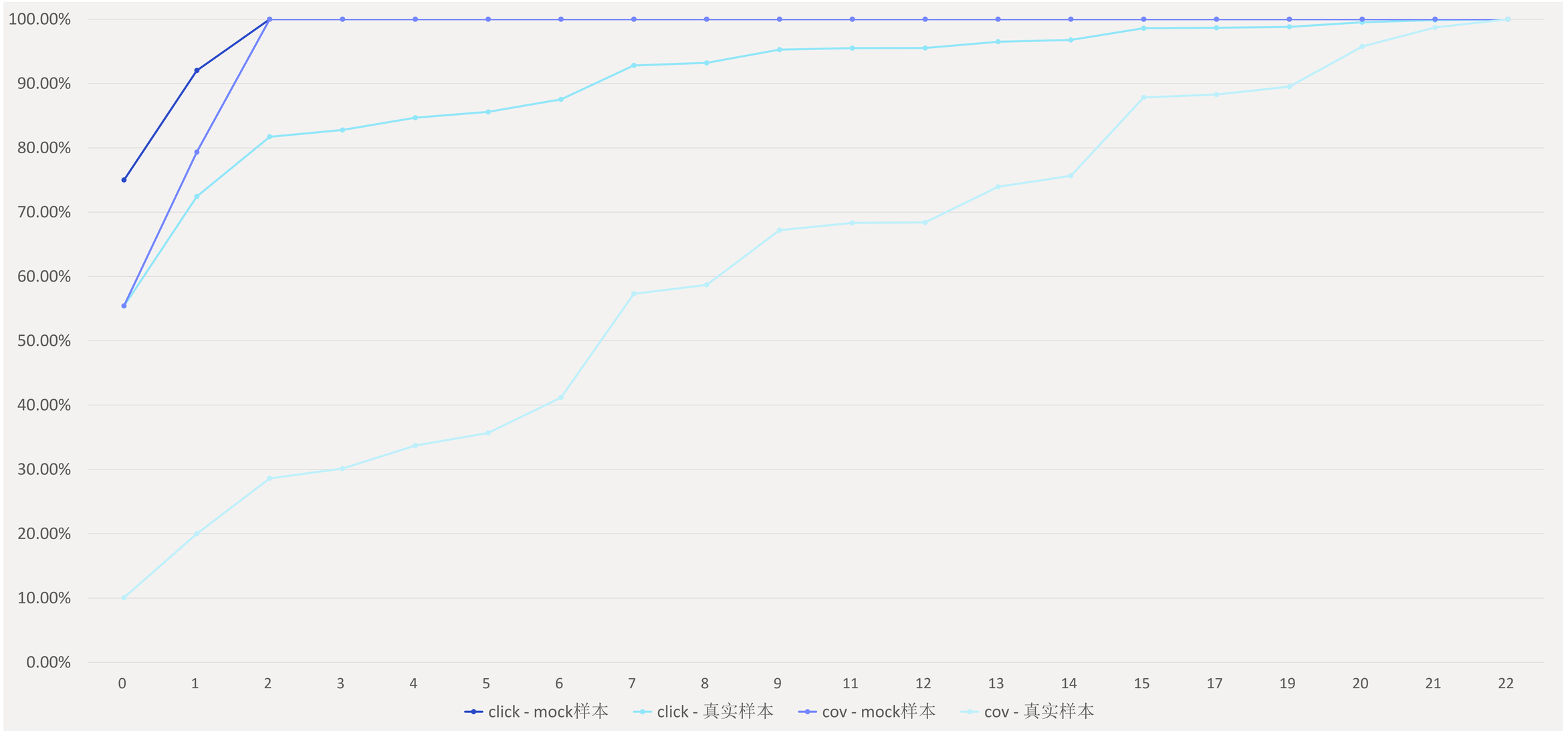
首次激活



横坐标：cvr分桶，0代表小于1%，1代表小于2%....
纵坐标：累计值/汇总值

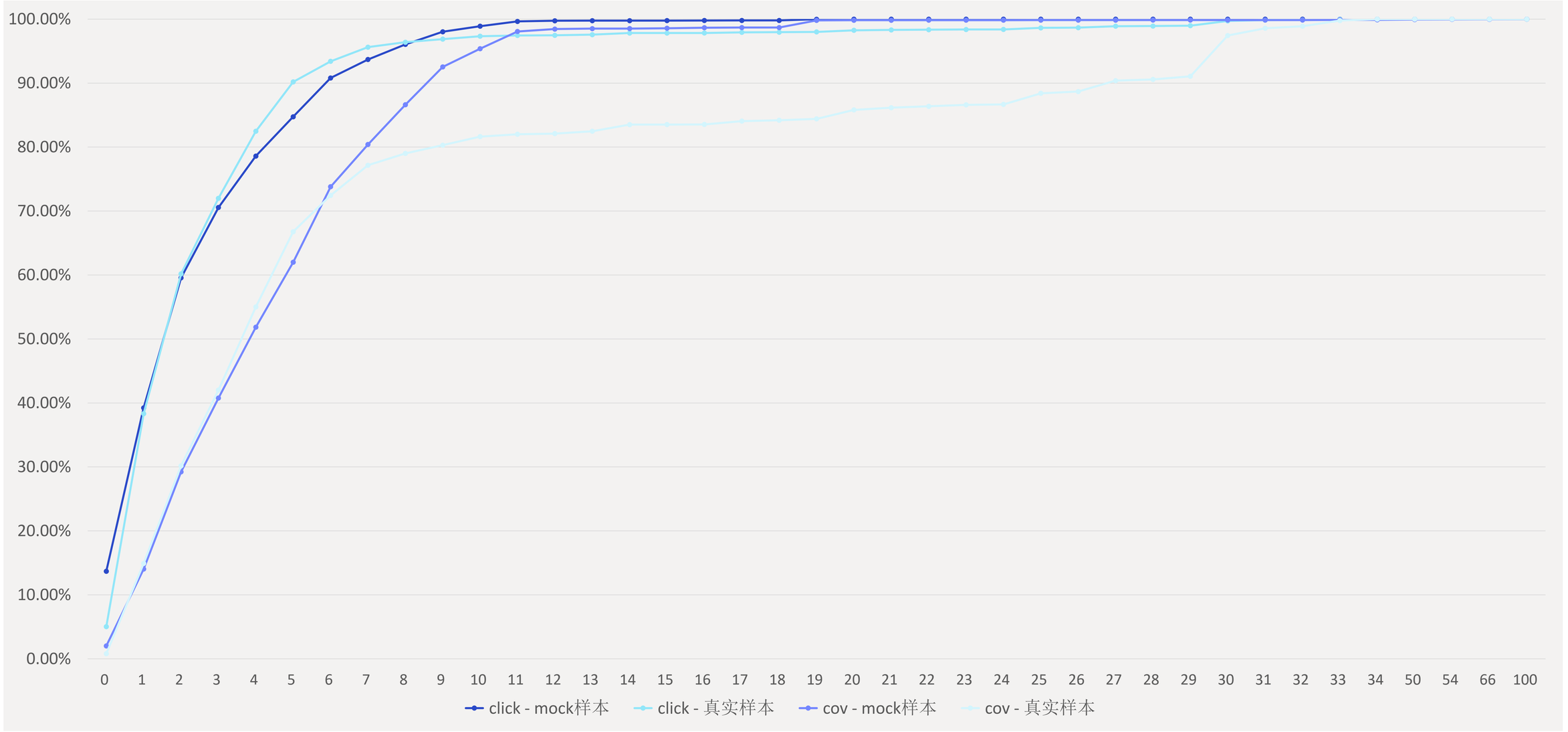
浏览器推荐页

表单提交



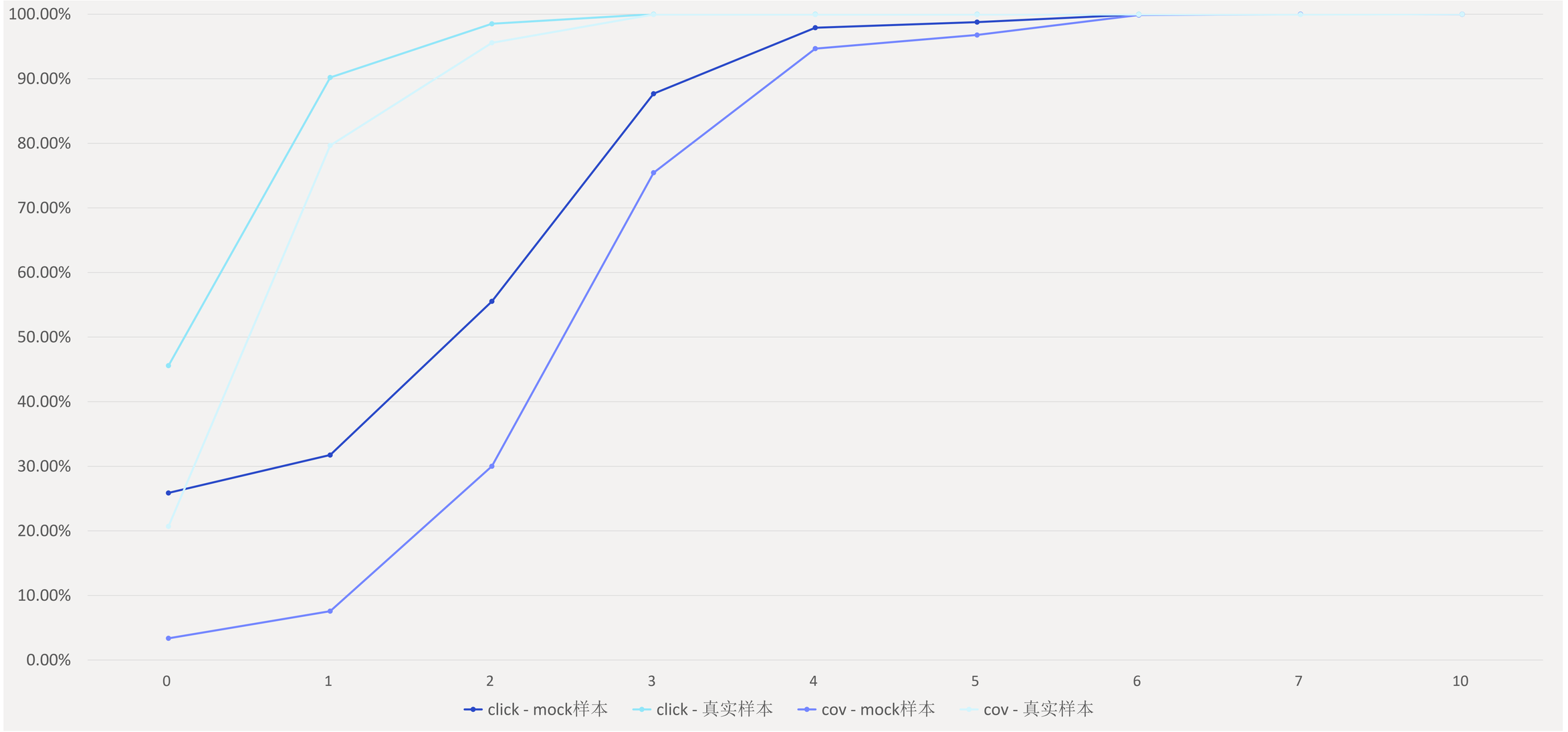
横坐标：cvr分桶，0代表小于1%，1代表小于2%....
纵坐标：累计值/汇总值

浏览器推荐页
自定义激活



横坐标：cvr分桶，0代表小于1%，1代表小于2%....
纵坐标：累计值/汇总值

浏览器推荐页
自定义注册



横坐标：cvr分桶，0代表小于1%，1代表小于2%....
纵坐标：累计值/汇总值

浏览器推荐页阈值

使用最近一个月的每周六的数据

■ **阈值设置的条件**：最近一个月，通过配置不同的阈值，使得归因结果中过滤ocpc点击/总ocpc点击，最大不超过1%，最好在0.5%以下

- **参考1**

		归因结果中过滤ocpc点击/总ocpc点击				归因结果中总过滤点击/总点击			
cvr阈值	浏览器推荐页	2020/2/29	2020/3/7	2020/3/14	2020/3/21	2020/2/29	2020/3/7	2020/3/14	2020/3/21
2.00%	下载	0.7%	2.4%	0.9%	1.2%	32.9%	43.5%	31.2%	35.5%
0.25%	游戏注册	0.0%	9.0%	4.8%	0.0%	1.5%	2.9%	2.4%	2.5%
1.00%	首次激活	1.4%	0.5%	1.2%	0.3%	37.4%	46.5%	35.4%	30.9%
0.10%	表单提交	1.5%	0.0%	0.0%	2.2%	1.4%	0.0%	0.0%	2.1%
0.30%	自定义激活	0.5%	0.3%	0.3%	0.5%	0.6%	0.6%	1.2%	2.2%

- **参考2**

		归因结果中过滤ocpc点击/总ocpc点击				归因结果中总过滤点击/总点击			
cvr阈值	**浏览器推荐页**	**2020/2/29**	**2020/3/7**	**2020/3/14**	**2020/3/21**	**2020/2/29**	**2020/3/7**	**2020/3/14**	**2020/3/21**
1.50%	**下载**	**0.4%**	**1.1%**	**0.4%**	**0.7%**	**32.3%**	**43.1%**	**30.9%**	**35.3%**
0.20%	游戏注册	0.0%	3.2%	0.0%	0.0%	1.3%	0.4%	1.1%	1.1%
0.80%	**首次激活**	**0.9%**	**0.4%**	**0.9%**	**0.3%**	**33.1%**	**42.1%**	**26.9%**	**27.2%**
0.05%	表单提交	0.0%	0.0%	0.0%	1.0%	0.0%	0.0%	0.0%	1.0%
0.35%	自定义激活	0.7%	0.5%	1.0%	0.8%	1.3%	1.0%	1.7%	2.8%
	自定义注册	其oCPC广告转化率最低，无需加cvr过滤条件							

cvr为点击日期，向后归因7天，后验

讨论

下一步、未来规划

问题：

- 1、是否在回流词表中加入 cvtype+adid 的 cvr数据，作为新过滤条件?

  是。

- 2、回流词表中每条 cvtype+adid 的 cvr计算口径?

  以回流词表时间分区为时间区间（分区内当天的点击日期），cvtype+adid 归因转化量/当天adid的点击数。

  3日回流词表使用3天的归因转化总量，7日回流词表使用7日归因转化总量。

规划中：

1、样本质量监控：

真实样本覆盖率

2、样本量增广：

Mock样本、样本偏移

新白名单过滤条件

阈值变更

	现状		变更为	
cvtype	1-3日加权平均回流率	1-3日变异系数	1-3日加权平均回流率	1-3日变异系数
下载	0.9	0.1	0.9	0.1
注册	0.9	0.1	**0.8**	0.1
首次激活	0.9	0.1	**0.8**	0.1
表单提交	0.9	0.1	0.9	0.1
自定义激活	0.9	0.1	0.9	0.1
自定义注册	0.9	0.1	0.9	0.1

实验结果：目前未出结果

THANK YOU.
谢谢。