基于金融组合优化方法的AdPages权重分配方法回测分析报告[[1]](#footnote-0)

摘 要：鉴于目前同一Campaign下不同adpages权重赋值通常为100，10等比较固定的数值，本报告结合金融资产组合优化中收益最大与单位风险下收益最大两种方法，尝试对不同adpages权重配比进行优化。通过对2019年以来数据库中有记录的所有Campaign进行回测分析，统计假设检验结果表明，相对于数据库中记录的权重配比，采用通过金融资产组合优化方法计算所得的权重配比，在99%置信水平下，胜率和超额收益均值都大于0。即比较确信，报告中的配置方法具有相对优势。且回测结果显示，保持budget、bid的原有设定，若采用报告中配置方法计算所得的权重，回测分析中Campaign的总收益会比采用数据库记录权重获得的收益高出148929美元。

本报告共包含四部分。第一部分对优化方法进行说明；第二部分举例具体说明配置与效果分析方法；第三部分对2019年后数据库中有记录Campaign的回测结果进行统计分析；第四部分是对回测结果与配置方法的进一步分析与说明。

1. 分析方法

观察到目前同一Campaign下不同adpages权重赋值通常为100，10等比较固定的数值，希望借鉴金融资产组合优化方法，尝试对不同adpages权重配比进行优化。

金融资产组合配置与adpages权重配比有相似之处。金融资产组合配置通过设定不同的配置目标，使所得组合能够获得相比基准组合更高或更稳健的收益。这里，配置目标即组合优化目标，可以有收益最大、风险最小、Sharpe比率或单位风险所获得的收益最大，风险平价等；基准组合通常为等权重组合或其它主观设定组合（如股票、债券的60/40组合等）。在金融资产组合配置一般计算方法中，需要输入不同资产的收益率序列。对比同一Campaign下不同adpages权重配比，我们自然希望所得收益更高，收益与Campaign对应的budget、bid及不同adpages权重配比皆有关，这里我们将问题局部化，即假设Campaign对应的budget、bid给定，只对不同adpages权重配比进行优化。某一adpages获得利润的条件为revenue - cost > 0，即rpi -cpc>0。类比金融资产，cpc可看作资产买入价，rpi为卖出价，因而(rpi -cpc) / cpc 可作为adpages的收益率。由此，可依据同一Campaign下不同adpages过去一段时间的收益率(rpi -cpc) / cpc序列，给出不同adpages的权重配比。

考虑广告业务的特征，选取收益最大与单位风险下收益最大作为组合优化策略。以，表示adpages的收益率，为某时点存在的adpages数量，表示不同adpages的权重，为个adpages某段时间的收益率均值向量，为个adpages某段时间收益率的方差-协方差矩阵。、分别为不同adpages组合的收益率与标准差，即包含个adpages的Campaign的收益率与标准差。收益率最大作为优化目标的配置模型为：



，.......................................... (1)

单位风险下收益率最大为优化目标的配置模型，考虑了资产收益率及其稳定性，模型表示如下：



，........................................... (2)

分别以、表示优化目标为收益率最大与单位风险下收益率最大所得的配置权重，则不同adpages的权重配比为与的加权和，即

， .............................................. (3)

其中，为权重。

1. 举 例

选取编号为351115035的Campaign为例，说明对具体Campaign下不同adpages权重配比优化及效果分析方法。该Campaign下adpages对应数据相对齐备，能较好阐释配比与效果分析方法。

2.1 权重配置优化方法

首先说明adpages权重配置优化方法。以2019年2月17日编号为351115035的Campaign下不同adpages权重配比为例。在2019年2月17这一日，已有交易的数据截至日期为2019年2月15日，在该日有三个不同的adpages，三个adpages的id分别为：1643265、1643266、1643267。这里选取3为回望窗口的长度默认值，若同一Campaign下不同adpages的个数大于3，则为保证上述计算权重的配置模型(1)、(2)均有解，回望窗口的长度即为adpages的个数。从而，选取2月15日及该日前2个交易日各adpages的收益率即 (rpi -cpc) / cpc序列，见下表1[[2]](#footnote-1)：

表1：三个adpages的收益率序列

|  |  |  |  |
| --- | --- | --- | --- |
|  | 1643265 | 1643266 | 1643267 |
| 2019/2/13 | 1.2279 | -0.0861 | 0.4043 |
| 2019/2/14 | 2.1052 | 0.3404 | 0.3079 |
| 2019/2/15 | 0.8281 | 0.3689 | 0.1552 |

至此，前文所述收益率分别为表1中的三列数值。由表1中数据及式(1)、(2)、(3)可求得2019年2月17日id为1643265、1643266、1643267的三个adpages的权重值为：

表2：单一交易日中三个adpages的配置权重值

|  |  |  |  |
| --- | --- | --- | --- |
|  | 1643265 | 1643266 | 1643267 |
| 2019/2/17 | 0.8830 | 0.0350 | 0.0820 |

这里，式(3)中参数，即权重分配中更偏重adpages收益率的高低。

以上求得了一日adpages的权重值，所有日期不同adpages的权重配置结果均可由上述方法得到[[3]](#footnote-2)。

2.2 配置效果分析方法

接下来将说明配置效果分析方法。2019年2月17日数据库中id为1643265、1643266、1643267的三个adpages的权重记录值为[[4]](#footnote-3)：

表3：单一交易日中三个adpages的记录权重值

|  |  |  |  |
| --- | --- | --- | --- |
|  | 1643265 | 1643266 | 1643267 |
| 2019/2/17 | 15 | 10 | 10 |

将表3中权重相对值进行归一化处理，使各权重之和为一，结果见下表4：

表4：单一交易日中三个adpages的记录权重归一化值

|  |  |  |  |
| --- | --- | --- | --- |
|  | 1643265 | 1643266 | 1643267 |
| 2019/2/17 | 15/35 | 10/35 | 10/35 |

而下表5中呈现了2019年2月17日id为1643265、1643266、1643267的三个adpages的收益率[[5]](#footnote-4)：

表5：单一交易日中三个adpages的收益率

|  |  |  |  |
| --- | --- | --- | --- |
|  | 1643265 | 1643266 | 1643267 |
| 2019/2/17 | 0.7171 | -0.1748 | 0.1258 |

由表5中收益率与表4中数据库记录权重值对应相乘再相加，可得到2月17日编号为351115035的Campaign的收益率为0.2933[[6]](#footnote-5)。由表2和表5，可得利用配置模型（1）~（3）计算所得权重对应的收益率为0.6374[[7]](#footnote-6)。而交易收益率可由当日三个adpages的revenue之和与cost之和相除得到，为0.2901[[8]](#footnote-7)。观察到真实收益率与由数据库记录权重值计算所得收益率略有差异，原因是各adpages的数据库记录权重值间比例关系不完全等于对应impression间比例关系。

显然由模型计算所得收益率与由数据库记录权重值所得收益率间更具可比性。2019年2月17日，配置模型（1）~（3）计算所得不同adpages权重的对应收益率0.6374高于数据库记录权重对应收益率0.2933。只要数据可得，所有交易日adpages的配置权重均可通过模型（1）~（3）计算得到，从而可得每个交易日Campaign的收益率。各交易日两收益率序列见下图1，其中“ret\_ratio\_real\_weight\_old”表示由数据库记录权重值所得收益率，“ret\_ratio\_weight\_alloc”表示由配置模型（1）~（3）计算所得权重对应的收益率。由图1，

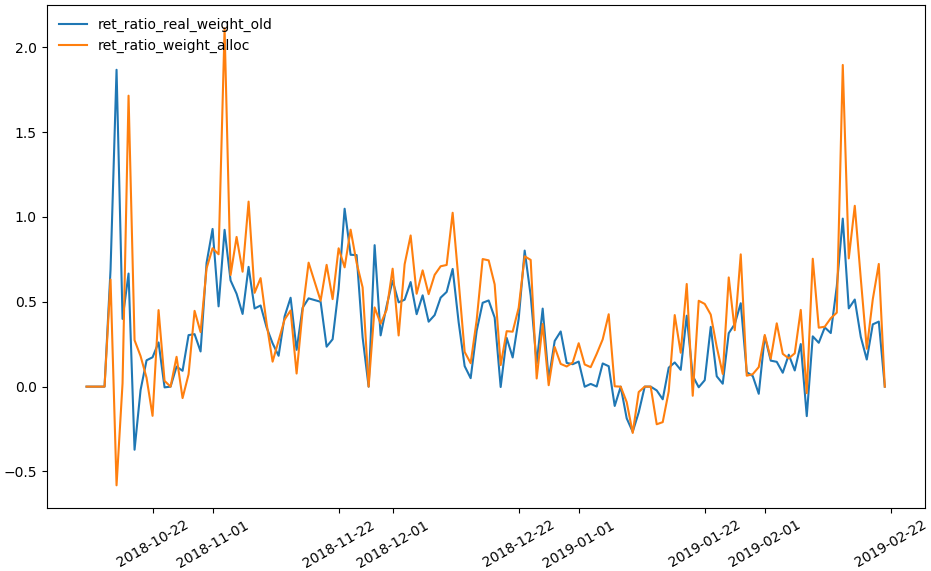


图1：各交易日由模型计算所得收益率与由数据库记录权重值所得收益率序列

在大多数交易日，配置模型（1）~（3）计算所得权重对应的收益率相对更高。将所有交易日中，配置模型（1）~（3）计算所得不同adpages权重的对应收益率高于数据库记录权重对应收益率的占比，与配置模型计算所得权重对应收益率低于数据库记录权重对应收益率的占比，二者作差，称所得差值为借鉴金融资产组合优化方法的配置模型的**胜率**。经统计，依据编号为351115035的Campaign的所有可得记录数据，所有交易日中，配置模型（1）~（3）计算所得不同adpages权重的对应收益率高于数据库记录权重对应收益率的占比为66.16%，配置模型计算所得权重对应收益率低于数据库记录权重对应收益率的占比为26.31%，胜率为39.85%。显然胜率越高意味着在更多交易日中，配置模型计算所得权重对应收益率更高，收益值也更高。

由收益率与对应交易日的成本cost相乘，可得当日收益值[[9]](#footnote-8)，各交易日两收益值序列见下图2，其中“ret\_real\_weight\_old”表示由数据库记录权重值所得收益，“ret\_weight\_alloc”表示由配置模型（1）~（3）计算所得权重对应的收益。

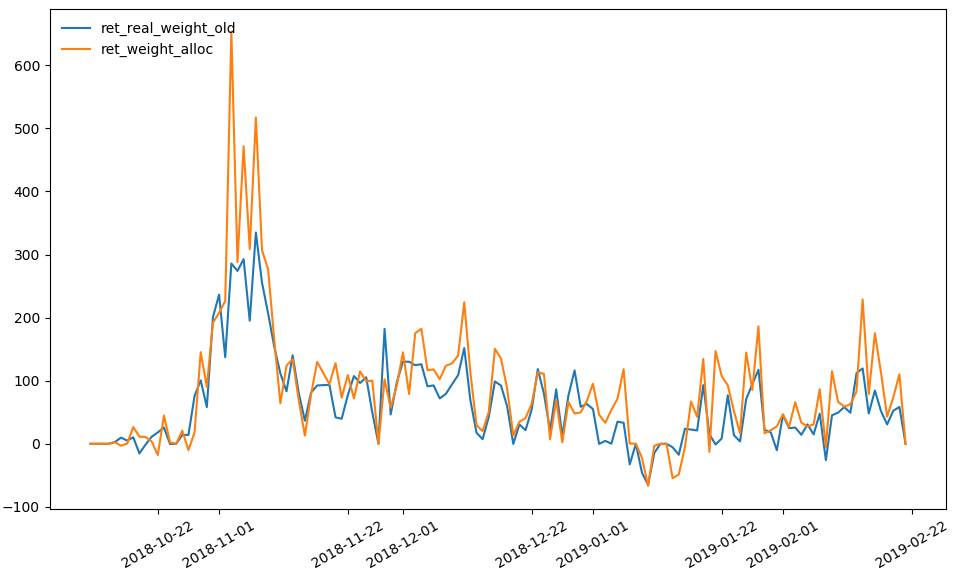


图2：各交易日由模型计算所得收益与由数据库记录权重值所得收益序列

将各交易日收益值累加，即为包含不同adpages的编号为351115035的Campaign总收益序列[[10]](#footnote-9)，见图3，其中“cum\_ret\_real\_weight\_old”表示由数据库记录权重值所得累计收益，“cum\_ret\_weight\_alloc”表示由配置模型（1）~（3）计算所得权重对应的累计收益。

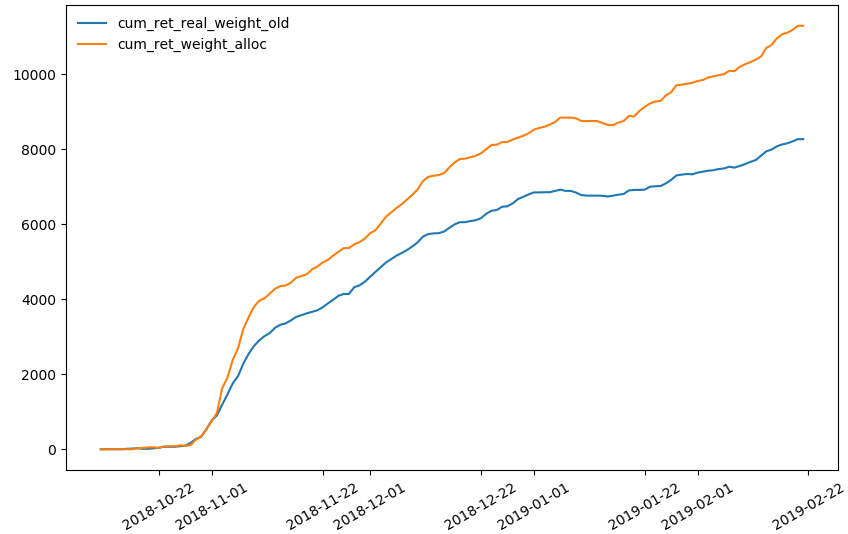


图3：由模型计算所得总收益与由数据库记录权重值所得总收益序列

观察图3，借鉴金融资产组合优化方法的配置模型（1）~（3）计算所得权重对应的累计收益明显高于由数据库记录权重所得累计收益，称两者差值为**超额收益**。截至2019年2月17日超额收益为2936美元，即对于编号为351115035的Campaign，在保持其budget和bid设置不变情况下，采用通过金融资产组合优化配置模型（1）~（3）计算所得的权重配置，相比数据库记录的权重配置，从数据有记录起始，截至2019年2月17日，总收益高出2936美元。

综上，将胜率与超额收益作为度量指标，对于编号为351115035的Campaign，借鉴金融资产组合优化方法的配置结果相对数据库记录的权重配置，具有优化效果。

1. 回测结果统计分析

对2019年后数据库[[11]](#footnote-10)中有记录的所有Campaign进行回测，除部分因数据不足，未能成功计算权重的Campaign外，共对近37500个Campaign成功进行了相应adpages的权重回测计算。这其中，部分Campaign在占比80%及以上较长时段数据不足或只有一个数据充足的有效adpage，此时“配置”没有实际意义，应从budget层面进行利润最大化设置。因此，这里着重分析配置有效的Campaign的优化情况，此类Campaign共计12210个。所有回测结果均记录在数据库adm\_files[[12]](#footnote-11)中。

下面对上述12241个Campaign进行分析，相关回测结果参见数据库中表格“compares\_statistics\_df”与“compares\_statistics\_df\_1”，其中“diff\_wl\_rate”列表示胜率，“diff\_cum\_crwa\_crrwo”列表示超额收益。

首先对胜率这一配置优化效果度量指标进行分析。下图4呈现了该指标的分布直方图，表6则为该指标的一系列统计值。

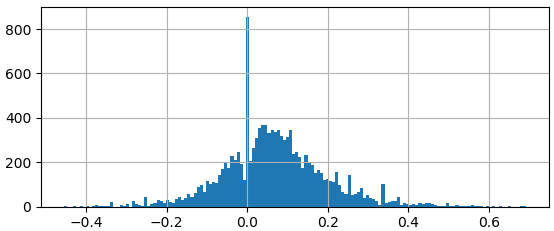


图4：胜率分布直方图

表6：胜率统计值

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 样本数量 | 均 值 | 方 差 | 最小值 | 25%分位 | 50%分位 | 75%分位 | 最大值 |
| 12210 | 0.0671 | 0.1294 | -0.4545 | 0.0000 | 0.0588 | 0.1373 | 0.6923 |

由图4，大部分Campaign的胜率大于0，经统计，这一比例为68.42%。此外，结合表6中最值与各分位数，可知胜率的取值相对于0是右偏的。进一步，采用t-检验，来验证胜率均值大于零这一原假设，由均值与方差及样本数量，可得相应t-统计值为57.27，显然在99%（甚至更高）置信水平下，胜率均值大于0。也就是说，平均来看，在更多交易日中，配置模型（1）~（3）计算所得不同adpages权重的对应收益率高于数据库记录权重对应收益率。

下面对度量指标超额收益进行分析。下图5呈现了该指标的分布直方图，表7则为该指标的一系列统计值。

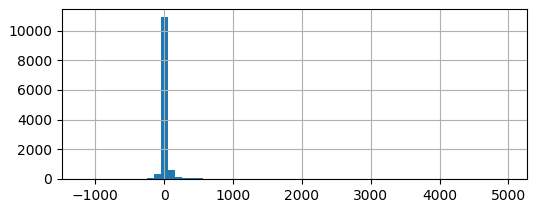


图5：超额收益分布直方图

表7：超额收益统计值

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 样本数量 | 均 值 | 方 差 | 最小值 | 25%分位 | 50%分位 | 75%分位 | 最大值 |
| 12210 | 12.20 | 109.58 | -1176.96 | -4.98 | 2.57 | 12.88 | 4963.79 |

同样，由图5，大部分Campaign的超额收益大于0，经统计，这一比例为60.80%。此外，结合表7中最值与各分位数，可知超额收益的取值相对于0是右偏的。进一步，采用t-检验，来验证超额收益均值大于零这一原假设，由均值与方差及样本数量，可得相应t-统计值为12.30，显然在99%（甚至更高）置信水平下，胜率均值大于0。也就是说，平均来看，配置模型（1）~（3）计算所得不同adpages权重的对应收益高于数据库记录权重对应收益。此外，对所有超额收益值求和，可得148929，即对此12210个Campaign，保持budget、bid的原有设定，若采用配置模型（1）~（3）计算所得的权重，将比采用数据库记录权重获得的收益高出148929美元。

为完整起见，对其余Campaign的回测结果也做简要分析。相关回测结果参见数据库中表格“compares\_statistics\_except\_df”与“compares\_statistics\_except\_df\_1”。图6、图7中分别呈现胜率与超额收益直方图，表8、表9分别给出两指标的统计值。

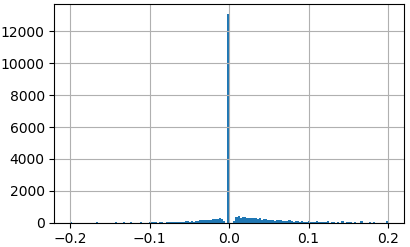
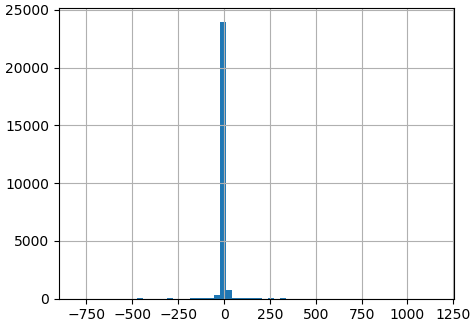
 

图6：胜率分布直方图（其余Campaign） 图7：超额收益分布直方图（其余Campaign）

表8：胜率统计值（其余Campaign）

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 样本数量 | 均 值 | 方 差 | 最小值 | 25%分位 | 50%分位 | 75%分位 | 最大值 |
| 25238 | 0.0107 | 0.0447 | -0.2000 | 0.0000 | 0.0000 | 0.0208 | 0.2000 |

表9：超额收益统计值（其余Campaign）

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 样本数量 | 均 值 | 方 差 | 最小值 | 25%分位 | 50%分位 | 75%分位 | 最大值 |
| 25238 | 0.40 | 18.69 | -803.00 | 0 | 0 | 1.0 | 1156 |

t-检验值分别为37.93、3.37，显然同样可认为在99%置信水平下，胜率和超额收益均值都大于0。特别地，将25238个超额收益值求和，值为9998美元。

综上，2019年后数据库中有记录的所有Campaign的回测结果显示，相对于数据库中记录的权重配比，采用通过金融资产组合优化配置模型（1）~（3）计算所得的权重配置，在99%置信水平下，胜率和超额收益均值都大于0。即比较确信，此优化配置模型具有相对优势。

1. 进一步分析与说明

4.1 配置模型表现不佳情形分析

下面对配置模型表现不佳的情形进行分析。

对于前面编号为351115035的Campaign，配置模型计算所得权重对应收益率低于数据库记录权重对应收益率的占比为26.31%。仔细分析这些交易日的情形，发现在这些交易日中不同adpages收益率的相对优势发生了突变或转折，而配置模型依据adpages收益率的历史表现进行权重设置，在此种情况下会赋予历史上收益率表现较好的adpage更大的权重，因而突变的发生会导致当日总收益较低。此外，在adpages创建初期，一般由于不能尽早找到收益率相对优势的规律，也会造成配置模型表现不佳。在adpages创建初期，产生配置模型所需的初始数据，可能由于初始资金量budget较小，若adpages权重设置不够均衡，会使得某些adpages的impression较少，从而可能造成click为0的情况，这种偶然性会导致不能尽早捕获adpages收益率的规律，因此在adpages初始创建时，尽可能为不同adpages设定较为均衡的权重配比，或有助于配置模型对不同adpages收益率信息的捕获。

此外，对其余超额收益较小的Campaign进行分析。以超额收益倒排前两名的编号为351062694与351028109的Campaign为例[[13]](#footnote-12)，经分析发现，Campaign包含的adpages较多，数据库中记录的权重配比在不同adpages间不断大幅跳转，且出现收益率持续较高的adpages权重配比值突然设置为0的情形，此类设置会影响收益率数据的表现，这些情况会使得配置模型不能找到adpages收益率表现的规律，配置结果不佳。

最后，如若Campaign下所有adpages的收益率均告负，则无论如何配置都会产生亏损，此时应置换adpages或从budget层面进行设置以使收益最大化。

4.2 可能的改进

当前配置模型包含一些参数，这些参数的取值均是结合业务特征的主观取值，未经优化，考虑相应时间成本，可在必要时再行优化。

此外，当前配置方法仅依据历史数据，当数据规律突然间发生巨大变化时，配置结果表现不佳。因此，若能对这些变化给出有效预测，则可采用预测数据替换历史数据，配置模型或可给出更优的配比结果。

另外，据了解业务员应用系统中有关于展示成本等的即时更新数据，该类数据提供了部分业务相关最新信息，或可考虑在某几个时间节点对此类数据进行采集，可能有助于把握adpages收益率的最新变化。

4.3 应用实施

本报告中的配置方法计算步骤相对简单。在某一交易日，最新交易数据到达后，可以开始计算Campaign下不同adpages的权重配比。计算步骤如下：

第一步，收集Campaign当前包含所有adpages近期时段的rpi与cpc数值，并计算出对应收益率，即准备数据；

第二步，根据不同adpages的收益率序列，计算配置权重值；

第三步，将归一化的权重值转化为相对比值（与系统设置模式匹配），作为即刻至下一交易日新数据到达前这一时段adpages的权重配比建议值。

本报告中配置方法，每日所需计算时间应该在可接受范围。根据回测运算耗时，可对计算时间做出估计。采用当前20个进程同步运算，所有Campaign回测耗时约24\*10=240小时，因中间有两次与服务器连接中断，因此240小时为耗时上限。这里对2019年后数据库中有记录的Campaign进行了回测运算，每一Campaign在数据库中记录的时长不一，保守估计平均为40个交易日，则每日全部运行中的Campaign计算一次所含adpages的权重，所需时长上限为240/40=6小时。若有更多运算资源，该时间可再缩短。

1. 回测与分析过程中涉及的程序与文档存放在服务器192.168.101.70上数据库adm\_files，以及文件夹/home/wumh/adm\_prediction中，以备查询。 [↑](#footnote-ref-0)
2. 该表格中数据可由表df\_adpages\_processed\_351115035.xls中数据得到（参见“return\_ratio”列） [↑](#footnote-ref-1)
3. 具体计算结果参见表df\_adpages\_processed\_351115035.xls中“weight\_alloc”列。需要说明的是，表格中“adjust\_date、adpage\_id，……”等列中数值由数据库aidata中ai\_weight20181126和ai\_weight两张表拼接所得，“return\_ratio”列由 (rpi -cpc) / cpc计算所得，“weight\_alloc”列中存在缺失值的原因是计算配置模型(1)、(2)所需数据不足。 [↑](#footnote-ref-2)
4. 参见表df\_adpages\_processed\_351115035.xls中“weight\_old”列。 [↑](#footnote-ref-3)
5. 参见表df\_adpages\_processed\_351115035.xls中“return\_ratio”列。 [↑](#footnote-ref-4)
6. 参见表camp\_ret\_ratio\_351115035.xls中“ret\_ratio\_real\_weight\_old”列。 [↑](#footnote-ref-5)
7. 参见表camp\_ret\_ratio\_351115035.xls中“ret\_ratio\_weight\_alloc”列。 [↑](#footnote-ref-6)
8. 参见表camp\_ret\_ratio\_351115035.xls中“ret\_ratio\_real”列。 [↑](#footnote-ref-7)
9. 参见表camp\_ret\_ratio\_351115035.xls中“ret\_real\_weight\_old”与“ret\_weight\_alloc”列。 [↑](#footnote-ref-8)
10. 参见表camp\_ret\_ratio\_351115035.xls中“cum\_ret\_real\_weight\_old”与“cum\_ret\_weight\_alloc”列。 [↑](#footnote-ref-9)
11. 服务器192.168.101.71上aidata库。 [↑](#footnote-ref-10)
12. 服务器192.168.101.70上adm\_files库。 [↑](#footnote-ref-11)
13. 分析数据参见表格df\_adpages\_processed\_351062694.xls和camp\_ret\_ratio\_351062694.xls，以及df\_adpages\_processed\_351028109.xls和camp\_ret\_ratio\_351028109.xls。 [↑](#footnote-ref-12)