

异动分析技术方案—异动归因之指标拆解

作者：伊璘、小椰

2022财年团队数据官暨FBI达人赛优秀案例

阿里数据 × 奇点学堂 数据学堂 FBI



数据学堂2022团队数据官暨FBI达人赛之优秀案例线下分享视频讲解

[视频：数据学堂 2022 团队数据官暨 FBI 达人赛之优秀案例线下分享]

异动分析技术分两部分，共四个章节，尽可能全面的探讨如何用机器学习模型做好异动分析。这里是第二部分的第一节异动分析技术方案—异动归因之指标拆解。

第一部分：[《异动分析技术方案—异动检测》](#)

第二部分 第一节：[《异动分析技术方案—异动归因之指标拆解》](#)（本篇）

第二部分 第二节：[《异动分析技术方案—异动归因之相关指标》](#)

第二部分 第三节：[《异动分析技术方案—异动归因之相关事件》](#)

综述：[《在ICBU，我们这样做指标异动分析的！》](#)

一、背景

前文第一部分异动检测在检测出异常之后，这一部分我们来谈一下归因，归因的方法有多种，这篇文章的重点是指标拆解，也是我们做业务分析时最常用到的方法。我们的目的是解放人力，将指标拆解实现自动化，一方面可以加快业务迭代速度，快速定位问题；另一方面可以对可能产生异动的维度进行全局量化，增强可比性，明确下一步的业务行动点的优先级。自动化异变归因的目的是为了尽快判断并抓住机遇，寻求以数据驱动作为灯塔指引业务航向。

二、目的

将目标指标定义为 Y ，波动为 $\Delta Y\% = \frac{Y^1 - Y^0}{Y^0}$ ，其中 Y^1 是当月的数据， Y^0 为上个月（同比/环比）的数据。

文章目的是为了研究组成 Y 的集合 $\{X\}$ 对于 $\Delta Y\%$ 的贡献：

$$\Delta Y\% = C_{X_1} + C_{X_2} + C_{X_3} + \dots$$

其中， C_{X_i} 表示指标（或维度） X_i 对于 $\Delta Y\%$ 的贡献度（contribution）。

另外，贡献的拆解算法是根据组合方式不同决定，集合 $\{X\}$ 组成 Y 的方式包括：

1. 加法 $Y = X_1 + X_2 + X_3$ (例, 各渠道uv加和)
2. 乘法 $Y = X_1 \cdot X_2 \cdot X_3$ (例, 已知cpc*ctr=rpm下, 分别算出cpc, ctr对于rpm的贡献)
3. 比率型指标 $Y = \frac{P}{S} = \frac{\sum p_{X_i}}{\sum s_{X_i}}$ (例, 各广告计划的cpf, 或者各个渠道的cpuv等)

下文按照组成指标的方式的不同, 分别陈述计算贡献的方式。

三、贡献率的拆解方法

3.1 加法拆解

已知 $Y = \sum_i X_i$

目标波动 $\Delta Y\% = \frac{Y^1 - Y^0}{Y^0}$

贡献等于 $C_{X_i} = \frac{X_i^1 - X_i^0}{Y^0} = \frac{\Delta X_i}{Y^0}$, 证明见附录1。

举例 针对绝对值指标的维度拆解都是加法拆解。绝对量指标的同比/环比变化, 就是各个分指标变化的加权求和, 例如访问uv总和等于各渠道uv加总, 那么总uv的变化下钻贡献率等于各渠道分别的变化除以上个月的总uv数。

3.2 乘法拆解

已知 $Y = \prod_i X_i$

目标波动 $\Delta Y\% = \frac{Y^1 - Y^0}{Y^0}$

贡献等于 $C_{X_i} = \frac{L(Y^1, Y^0) \cdot \ln\left(\frac{X_i^1}{X_i^0}\right)}{Y^0}$,

其中 X^1 是当月的数据, X^0 为上个月 (同比/环比) 的数据, $L(Y^1, Y^0)$ 为平均对数权重:

$$L(Y^1, Y^0) = \frac{Y^1 - Y^0}{\ln(Y^1) - \ln(Y^0)} = \frac{\Delta Y}{\ln(Y^1) - \ln(Y^0)}$$

证明见附录1。

举例 漏斗模型, 借助用户动线, 可以把指标拆解。以全站ipv为例, 其变动涉及站外用户引流、uv-d转化、detail页用户人均pv量, 分别对应了用户增长、搜推场景承接以及私域用户活跃度等业务域或用户行为指标。借此对全站ipv的构成链路进行静态乘法拆解:

$$ipv = uv \times \frac{duv}{uv} \times \frac{ipv}{duv} = uv \times (uv - d转化率) \times (人均ipv)$$

备注1:通常我们看的是l-d转化率, 由于本例是站在全站视角去拆解的, 链路较粗, 看uv-d转化率。

同时, 我们可以计算各乘积因子对目标指标变化的贡献率 C_i , 衡量3个指标的重要性:

$$C_{uv} = L(ipv^1, ipv^0) \cdot \ln\left(\frac{uv^1}{uv^0}\right)$$

$$C_{uv-d \text{ rate}} = L(ipv^1, ipv^0) \cdot \ln\left(\frac{uv-d \text{ rate}^1}{uv-d \text{ rate}^0}\right)$$

$$C_{ipv \text{ per } duv} = L(ipv^1, ipv^0) \cdot \ln\left(\frac{ipv \text{ per } duv^1}{ipv \text{ per } duv^0}\right)$$

$$L(ipv^1, ipv^0) \text{ 是平均对数权重, } L(ipv^1, ipv^0) = \frac{ipv^1 - ipv^0}{\ln(ipv^1) - \ln(ipv^0)} = \frac{\Delta ipv}{\ln(ipv^1) - \ln(ipv^0)}$$

举例2 广告p4p模型

$$\begin{aligned} cost &= pv \cdot \frac{adpv}{pv} \cdot \frac{imps}{adpv} \cdot \frac{click}{imps} \cdot \frac{cost}{click} \\ &= pv \cdot (cov \cdot asn) \cdot ctr \cdot cpc \end{aligned}$$

其中，pv代表平台流量规模和质量，cov*asn代表买卖双方匹配效率，ctr同时由买家质量、卖家供给和匹配质量决定，cpc与卖家消耗意愿和平台机制相关。

下图是pv, cov, asn, ctr, cpc对于消耗cost的贡献情况，X轴是以周为单位的时间序列，Y轴的折线图是消耗的周环比（增幅与上周相比），堆叠柱状图的颜色代表不同指标的贡献 C_{X_i} ，大于0代表是正向贡献，小于0是负向贡献，正负加起来的总和正是消耗的周环比 $\Delta Y\%$ 。



图一

3.3 比率型指标拆解

$$\text{已知 } Y = \frac{S}{P} = \frac{\sum_i s_{X_i}}{\sum_i p_{X_i}}, \quad P_i = \frac{p_{X_i}}{\sum_i p_{X_i}}, \quad S_i = \frac{s_{X_i}}{\sum_i s_{X_i}}, \quad Y_i = \frac{s_{X_i}}{p_{X_i}}$$

当分析比率指标进行维度下钻，分项对整体的贡献，受两个因素影响

- 分项的相对数指标波动贡献，即当期与基期的分项规模一致时，分项指标带来的变化：

分项的指标波动贡献 A_{X_i} = 指标同比变化值 * 上期基数占比

- 分项的结构变化，即当期与基期分项规模变化部分的指标变化：

分项的结构变化 B_{X_i} = 占比同比变化值 * (分项本期指标 - 整体上期指标)

$A_{X_i} = (Y_i^1 - Y_i^0) \times P_i^0$ ，其中 Y_i^1 是当月的数据， Y_i^0 为上个月（同比/环比）的数据。

$$B_{X_i} = (P_i^1 - P_i^0) \times (Y_i^1 - Y_i^0)$$

$$C_{X_i} = \frac{A_{X_i} + B_{X_i}}{Y_i^0}$$

举例 以uv-d转化率为例， $rate = \frac{duv}{uv}$ ，流量渠道大类可分为付费、免费、自然、其他，每种渠道的uv-d转化率为 $rate_i = \frac{duv_i}{uv_i}$ ，各渠道的duv占比用 $DUV_i = \frac{duv_i}{duv}$ 表示，各渠道uv占比用 $UV_i = \frac{uv_i}{uv}$ ，如果uv-d转化率同比下跌，我们想定位出哪个渠道出现了问题；各渠道的贡献 是怎么样计算为：

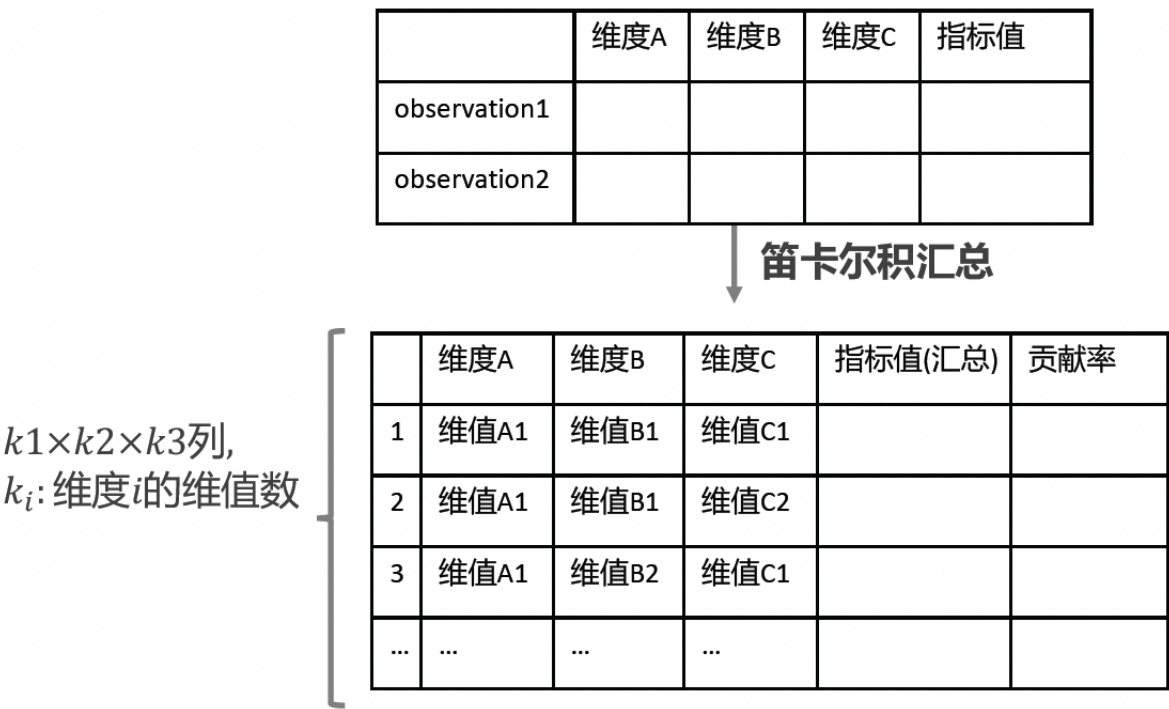
$$A_i = (rate_i^1 - rate_i^0) \cdot uv_i^0 = \Delta rate_i \cdot uv_i^0$$

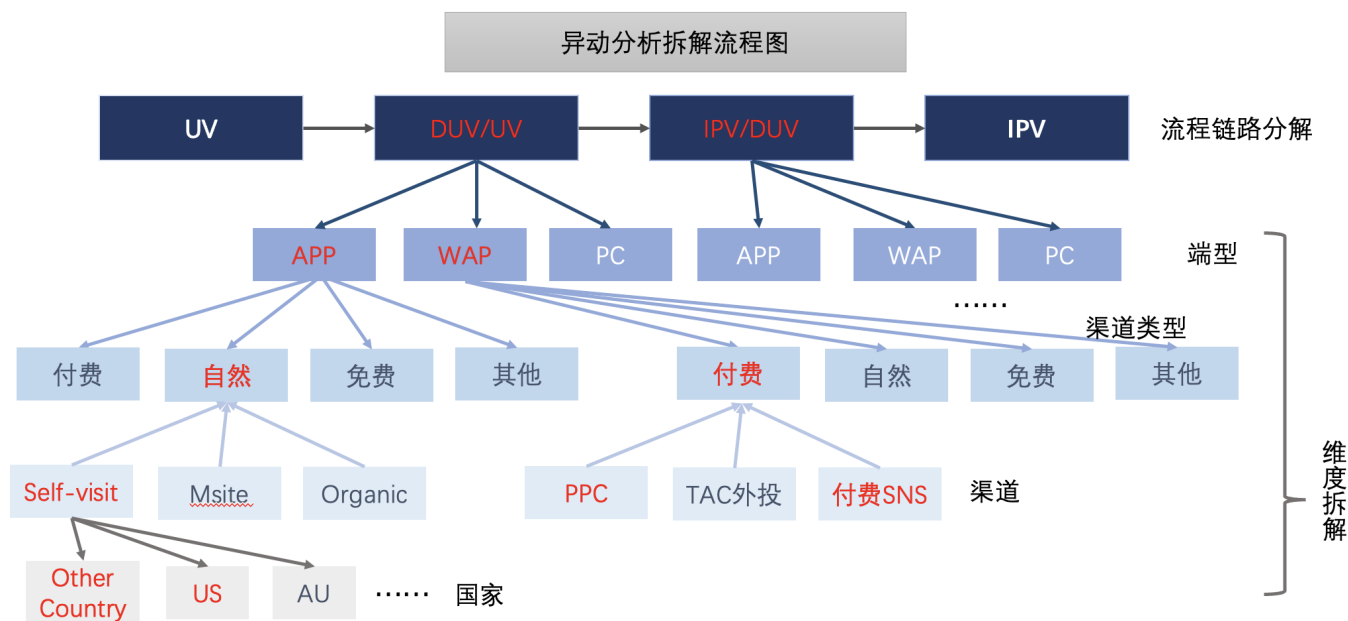
$$B_i = (uv_i^1 - uv_i^0) \cdot (rate_i^1 - rate_i^0) = \Delta uv_i \cdot (rate_i^1 - rate_i^0)$$

$$C_i = \frac{A_i + B_i}{rate_i^0}$$

3.4 实例应用

根据上文提到的不同指标的计算方法，支持全类型指标下钻求贡献的场景， 可根据先验业务输入搭建多层的归因逻辑模型， 层层下钻， 最终将指标波动定位。





图三 逐层下钻

红色表示的链路指标或维度代表对总值下跌贡献率较大，经过一层一层的拆解定位到app端自然流量中转化的降低导致总转化下降。

基于10月以来的流量跨端调控以及流量预算减投的业务背景，我们现将对ipv贡献最大的uv/duv根据端型、流量渠道类型、流量渠道、国家四个维度进行贡献率拆解。

在本实例中，通过本文对贡献率拆解方法与业务人工看数得到的问题定位基本一致，该方法可以实现异动贡献率量化与提效的目的，具体核心结论如下：

结论一（第一层拆解）ipv下降主要影响因素是uv-d转化率下降。

结论二（第二层拆解）uv-d比率的下降主要由APP端与WAP端导致，两种端型分别造成总比率约一半的下跌

结论三（第三/四层拆解）其中APP端的自然流量（self-visit）和wap端的付费流量（PPC、付费sns）是uv-d总比率下跌的主要贡献维度。

结论四（第五层拆解）APP端self-visit中美国与非核心17国uv-d比率下跌严重。

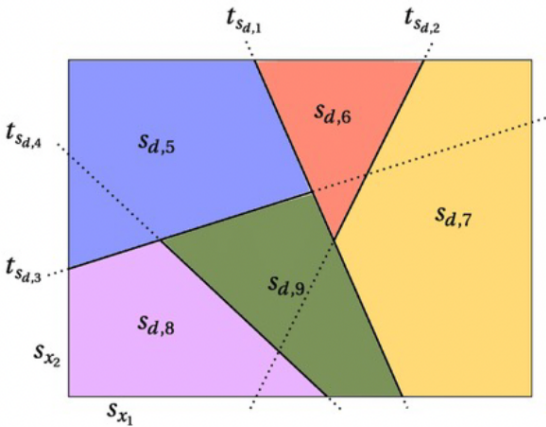
通过建立多层归因下钻维度模型，用自动化的方式层层剥析，从而能尽求完善且正确的归因到某个维度，从而节省人力，提高准确和科学性。

四、多层下钻归因方案一决策树

本节重点在拆解求出贡献率之后，如何探查异动。我们已经把不同维度下，每个维值的贡献率求出，下一步的目的是求出贡献最大（有异动的）的维度维值组合，测三种拆解方案，包括逐层下钻（同上文3.4的下钻方式）、多层同步下钻、决策树模型，发现决策树模型效果最好。这里决策树输入为不同的维值组合，输出为贡献率，做的是回归预测。

主要做法是求贡献率的方差（近似熵的概念），找到信息增益最高切割方法。这里自然而然想到决策树模型，通过贪心算法，切割数据空间，找到贡献率绝对值最高的维度组合空间。图四长方形整体表示数据空间，

s, t 表示两个维度，其下角标表示维度下的维值。下图具象的看出通过不同维值的组合，把数据空间切割成不同块，用不同的颜色代表。



图四

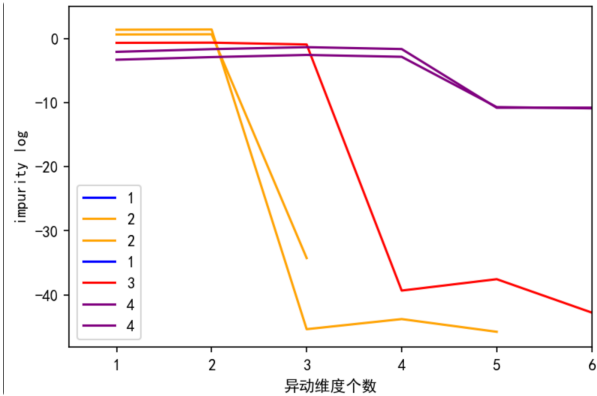
4.1 剪枝

决策树存在过拟合的问题， 为了解决这个问题，我们决定了剪枝的方法，采用后剪枝（Post-pruning）。后剪枝就是先把整颗决策树构造完毕，然后自底向上的对非叶结点进行考察，若将该结点对应的子树换为叶结点能够带来泛华性能的提升，则把该子树替换为叶结点。

后剪枝的方法包括：REP–错误率降低剪枝， PEP–悲观剪枝， CCP–代价复杂度剪枝， MEP–最小错误剪枝。我们借鉴了CCP—代价复杂度的方法。 选择节点表面误差率增益值大的层级的非叶子节点，删除该非叶子节点的左右子节点，若有多个非叶子节点的表面误差率增益值相同小，则选择非叶子节点中子节点数最多的非叶子节点进行剪枝。这个算法的参数为 $\alpha \geq 0$, 表示算法的复杂度:

$$\alpha_{eff}(t) = \frac{R(t)-R(T_t)}{|T|-1}$$

其中， $R(t)$ 表示的是结点 t 的方差（近似有权重的熵的概念：impurity，下文都简称为熵）， $R(T_t)$ 是结点 t 的子树 T_t 的熵的总和， $|T|$ 为决策树结点个数 。 α_{eff} 高，表示结点 t 往下分的信息增益高。图五表示异动维数的个数与决策树层结点熵的平均数的关系：以黄线为例， 当异动的维数为2时，决策树在第二层的熵最高，从第二层往后，再往下分熵越小，信息增益少，过拟合明显。从折线明显看到，熵的拐点在第二层，决策树最大深度等于2。我们从图五的事例启发，按照CPP的方法， 找跃层增益较大的“拐点”，找到合适的 α_{eff} 进行剪枝。



图五 异动维数的个数与结点熵的关系

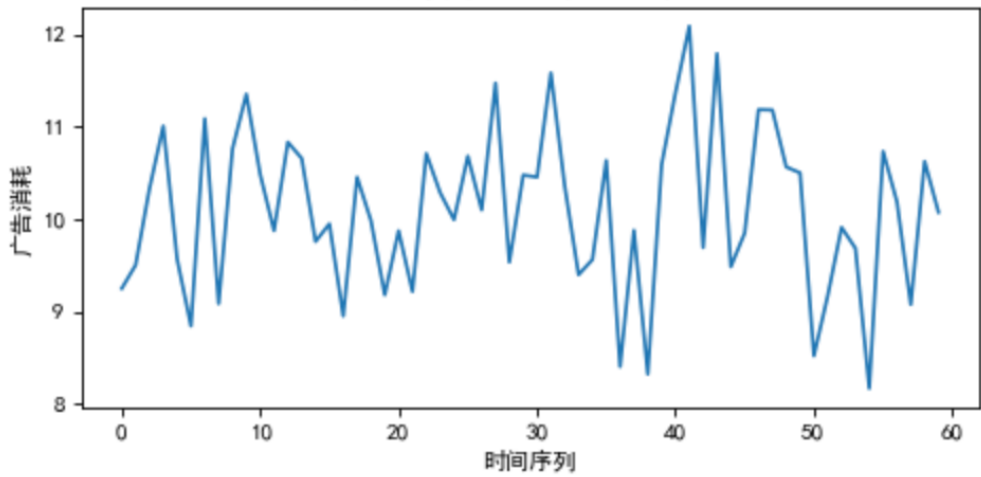
五、模型表现

5.1 模拟数据

我们模拟的维度和维值如下，共4个维度（两两独立），涉及维值共40个，4个维度维值组合（笛卡尔积 $3123 \times 4 = 744$ ）共744个。模拟的时间对比为月环比，模拟指标为广告消耗。

维度字段	country_cn_name	is_free	terminal_type_cd	imps_cnt_bins
维度	国家	渠道	端型	曝光档位
维值数	31	2	3	4
举例	马来西亚	免费	WAP	(-1.0, 0.0]

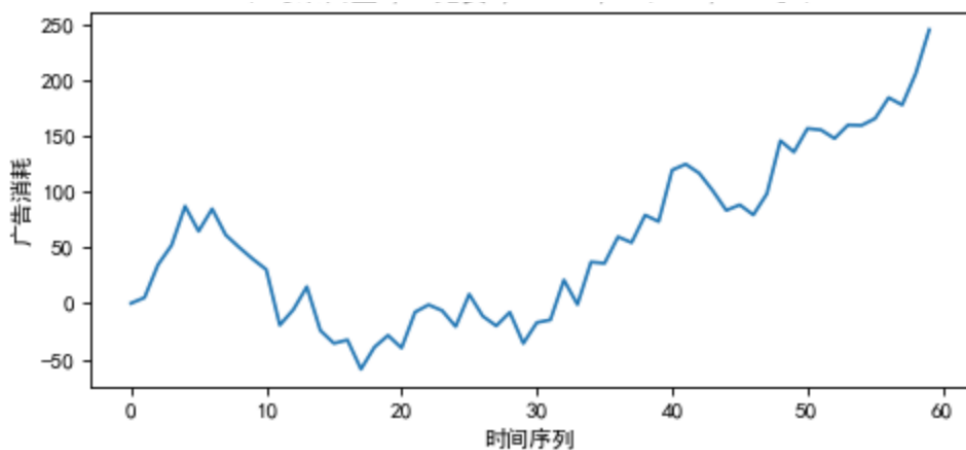
无异动数据：用白噪音 $x_t, x_{t+1}, x_{t+2} \dots \sim \mathcal{N}(10, 1)$ 模拟无异动的维度组合的时间序列，见图六



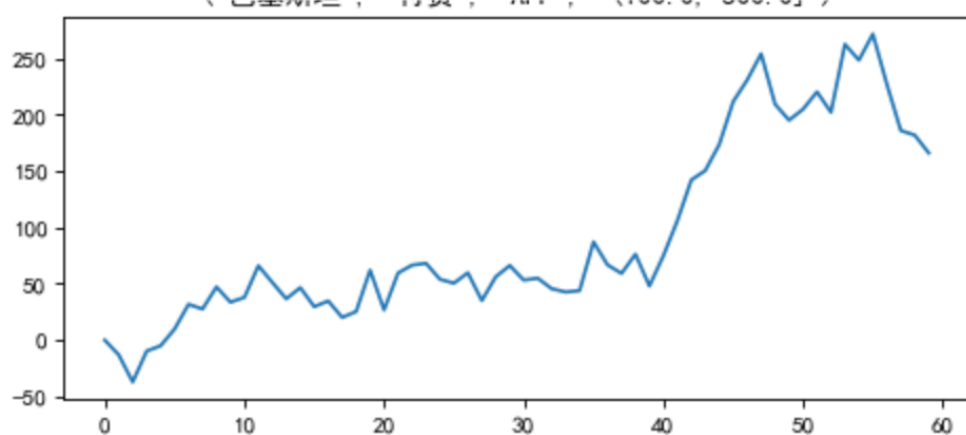
图六 无异动的时间序列

有异动数据：用随机游走的累积和来模拟异动，公式如下，见图七

$$x_0 \sim \mathcal{N}(3, 20), x_1 = x_0 + \mathcal{N}(3, 20), x_2 = x_1 + \mathcal{N}(3, 20) \dots$$



图七 (a): 有异动的时序列1



图七 (b): 有异动的时序列2

5.2 模型评估

在上图四个维度（国家、渠道、端型、曝光档位），指定特定的维度和维值在3月有异动，通过决策树模型，测试是否找到正确异动点。

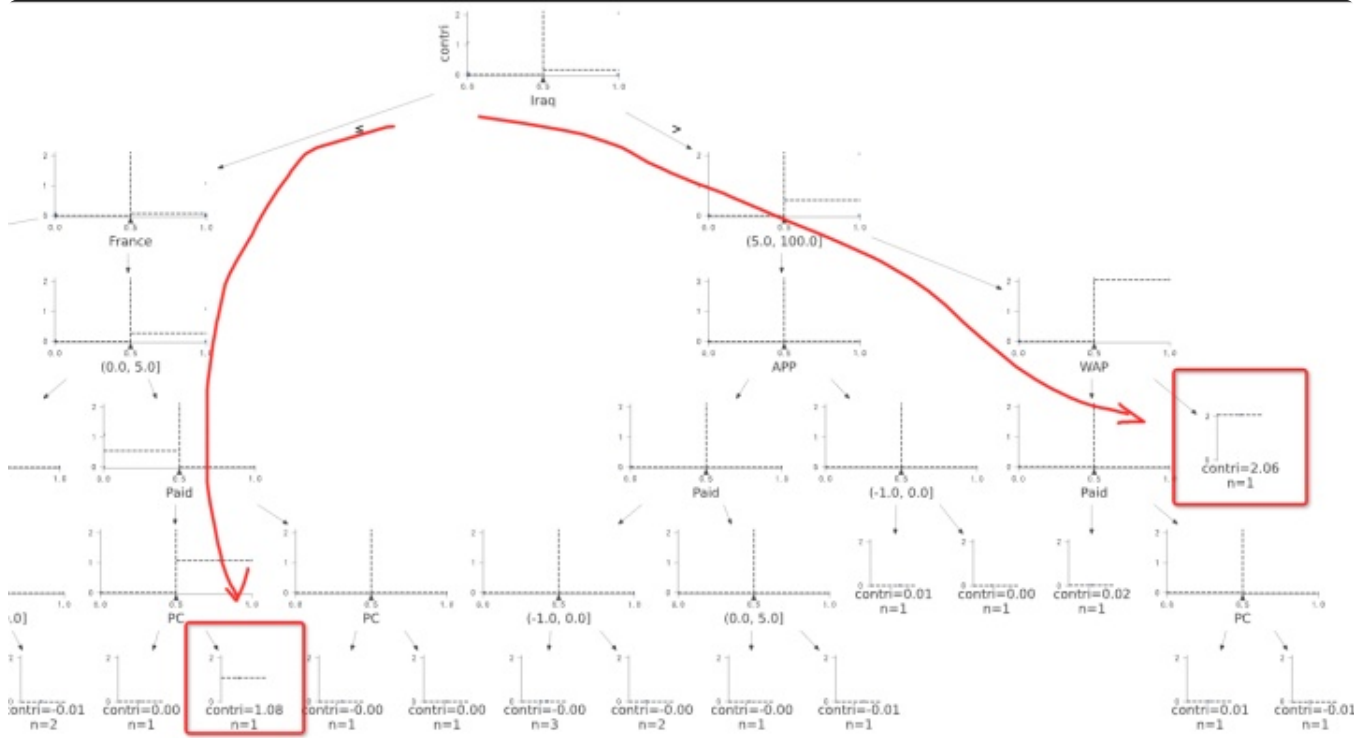
5.2.1 例一：异动维度在两处

异动维值组合：

a. 国家=伊拉克， 渠道=免费， 端型='WAP'， 曝光档位=[5:100]

b. 国家=法国， 渠道=免费， 端型='PC'， 曝光档位=[0:5]

将贡献度算出，数据输入决策树模型，结果见图八，可以看出决策树精确的找到异动的数据（共精确找到7个维值，共8个），且这两组标红数据对于异动的贡献绝对值最大。我们自定义树结构找父节点的方法，自动剪掉冗余分支，只截取重点枝干呈现。



图八：决策树结果呈现

特征重要性也符合预期：

国家=Iraq : 特征重要性=1.752

曝光档位=(5.0, 100.0]: 1.304

端型=WAP : 1.156

国家=France : 0.458

曝光档位=(0.0, 5.0]: 0.262

端型=PC : 0.221

渠道=Paid : 0.187

国家=澳大利亚 : 0.001

国家=西班牙 : 0.001

5.2.2 例二：异动维度在一处，只异动一个维度在付费上

异动维值组合：a. 渠道=付费

通过剪枝，模型成功找到一维信息，避免提供太多噪音令用户混淆。

5.2.3 更多

表一包含更多维值组合案例，以及模型表现，包括F1-score，模型输出的结果，和特征重要性。已探索11个案例，平均F1-score达到91.9%。

下面的数量是异动的维值个数：

Positive Class

True Positive (TP) 34

False Negative (FN) 6

False Positive (FP) 0

True Negative (TN) 0

Precision = 34 / (34 + 0) = 100%

Recall = 34 / (34 + 6) = 85%

F-1 Score Overall = 91.9%

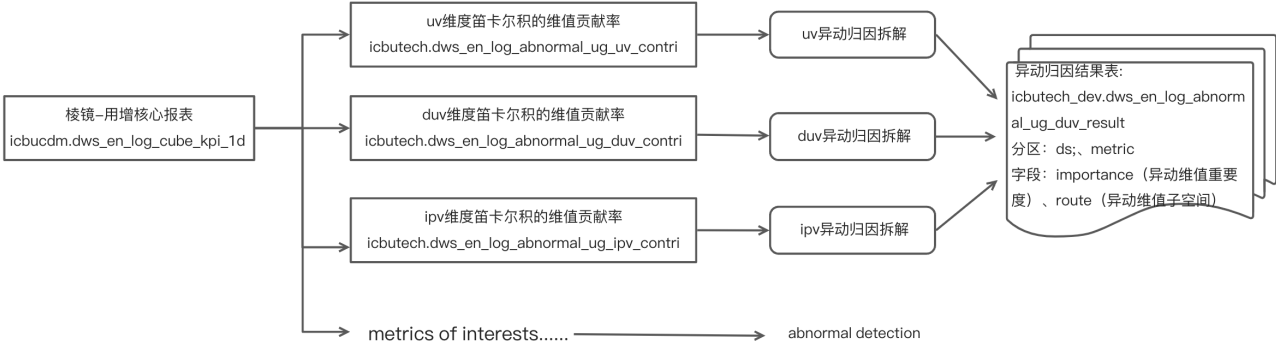
模拟case考虑的主要是可能存在异动的真实情况：1. 某个PID数据录入异常，会影响单维度的异动（仅那个PID的数据）。2. 某个渠道且某个端型的减投，会影响多个维度组合的异动。由于指标异动涉及的业务繁杂，不同团队在不同方向的优化，影响到不同的维值组合。

数量	异动维值组合 (label/y_true)	F1-Score	模型结果1: 模型输出（结果可视化展现，自动剪掉冗余分支，只截取贡献高的枝干呈现）	模型结果2:特征重要性
1	APP	100%	APP ->贡献率为 0.58	APP: 0.790
2	法国	100%	法国 ->贡献率为 0.84	法国 : 0.542;
3	付费	100%	付费 ->贡献率为 0.43	付费 : 0.570;
4	法国&付费	100%	法国 -> 付费 -> 贡献率为 0.81	法国 : 1.988; 付费 : 1.019;
5	法国&APP	100%	法国 -> APP -> 贡献率为 0.84	法国 : 1.895; APP: 1.360
6	APP&付费	100%	APP -> 付费 -> 贡献率为 0.41	APP: 1.591 付费 : 1.203;
7	APP&付费&法国	100%	法国 -> APP -> 付费 ->贡献率为 0.88	法国 : 1.944; APP: 1.374; 付费 : 1.047;
8	APP&付费&法国&(5.0, 100.0]	100%	法国 -> (5.0, 100.0] -> APP -> 付费 ->贡献率为 0.90	法国 : 1.836; APP: 1.469; (5.0, 100.0]: 1.369; 付费 : 1.302;
9	APP&付费&法国&(5.0, 100.0] or PC&免费&伊拉克	83%	非伊拉克 -> 法国 -> (5.0, 100.0] -> 贡献率为 0.14 伊拉克 -> PC -> 非付费 ->贡献率为 0.87	伊拉克: 1.610; PC : 1.027; 付费 : 0.772; 法国 : 0.062; (5.0, 100.0]: 0.052;
10	APP&付费&法国 or PC&免费&伊拉克&(5.0, 100.0]	83%	非法国 -> 伊拉克 -> (5.0, 100.0] -> 贡献率为 0.14 法国 -> APP -> 非免费 ->贡献率为 0.87	法国 : 1.559; APP: 1.101; 免费 : 0.839; 伊拉克: 0.064; (5.0, 100.0]: 0.058; PC : 0.000
11	APP&付费&法国 or PC&免费&伊拉克&(5.0, 100.0] or 美国&免费&APP	88%	非美国-> 非法国-> 伊拉克-> (5.0, 100.0]->贡献率为 0.14 非美国-> 法国-> APP-> 付费->贡献率为 0.84 美国-> APP-> 非付费->贡献率为 0.84	APP : 1.218 美国 : 0.936 付费 : 0.907 法国 : 0.863 伊拉克 : 0.035 (5.0, 100.0]: 0.033 土耳其 : 0.000

六、真实案例实践：棱镜-用增线上每日调度demo

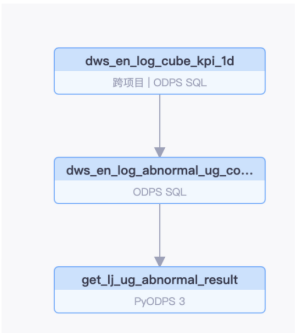
棱镜-用户增长域的核心指标的异动归因是当前以及未来异动归因最常用的场景之一，在象数异动归因产品正式上线前，对棱镜上用增的指标观测起来，对截止每日的当月指标日均值的年同比变动进行观测和维度拆解，定位贡献率最大的维值组合。

6.1 调度结构



图九 调度结构

odps调度节点:



图十 调度结点样式

6.2 效果展示

read icbutech_dev.dws_en_log_abnormal_ug_duv_result ;

结果[1]	结果[2]	结果[3]	结果[5]	结果[6]	查看MaxCompute队列
A	B	C	D	E	
total_diff_perc	importance	route	ds	metric	
-0.2761748196127744	其他 : 1.414; WAP: 1.333; pay: 0.731; 付费SNS: 0.113; Organic_Direct: 0.02	非WAP --> 非self_visit --> Facebook --> 其他 --> 贡献率为 -1.37; 非WAP --> s20220220	duv		
-0.2751871242279393	其他 : 1.413; WAP: 1.331; pay: 0.729; 付费SNS: 0.115; Organic_Direct: 0.02	非WAP --> 非self_visit --> Facebook --> 其他 --> 贡献率为 -1.39; 非WAP --> s20220221	duv		
-0.274543394126916	其他 : 1.411; WAP: 1.331; pay: 0.732; 付费SNS: 0.113; Organic_Direct: 0.02	非WAP --> 非self_visit --> Facebook --> 其他 --> 贡献率为 -1.41; 非WAP --> s20220222	duv		
-0.2767366453391990	其他 : 1.410; WAP: 1.332; pay: 0.734; 付费SNS: 0.111; Organic_Direct: 0.02	非WAP --> 非self_visit --> Facebook --> 其他 --> 贡献率为 -1.42; 非WAP --> s20220223	duv		
-0.2828492438226035	其他 : 1.408; WAP: 1.317; pay: 0.721; 付费SNS: 0.122; Organic_Direct: 0.02	非WAP --> 非self_visit --> Facebook --> 其他 --> 贡献率为 -1.42; 非WAP --> s20220224	duv		
-0.29056590786734	其他 : 1.407; WAP: 1.310; pay: 0.715; 付费SNS: 0.129; Organic_Direct: 0.02	非WAP --> 非self_visit --> Facebook --> 其他 --> 贡献率为 -1.42; 非WAP --> s20220225	duv		
-0.2994060990149597	其他 : 1.407; WAP: 1.306; pay: 0.712; 付费SNS: 0.132; Organic_Direct: 0.02	非WAP --> 非self_visit --> Facebook --> 其他 --> 贡献率为 -1.43; 非WAP --> s20220226	duv		
-0.3089647188934047	其他 : 1.408; WAP: 1.311; pay: 0.714; 付费SNS: 0.128; Organic_Direct: 0.02	非WAP --> 非self_visit --> Facebook --> 其他 --> 贡献率为 -1.42; 非WAP --> s20220227	duv		
-0.3184764394229223	其他 : 1.407; WAP: 1.314; pay: 0.714; 付费SNS: 0.125; Organic_Direct: 0.02	非WAP --> 非self_visit --> Facebook --> 其他 --> 贡献率为 -1.42; 非WAP --> s20220228	duv		
-0.4969389305899525	其他 : 1.353; WAP: 0.921; 付费SNS: 0.663; PPC: 0.202; self_visit: 0.037; Fac	非WAP --> 非self_visit --> Facebook --> 其他 --> 贡献率为 -2.09; 非WAP --> s20220301	duv		
-0.4848896004123816	其他 : 1.423; WAP: 1.101; 付费SNS: 0.630; PPC: 0.232; SEO: 0.035; self_vis	非WAP --> 非self_visit --> Facebook --> 其他 --> 贡献率为 -2.03; 非WAP --> s20220302	duv		
-0.500443256055887	其他 : 1.361; WAP: 1.073; 付费SNS: 0.347; pay: 0.174; self_visit: 0.031; Fac	非WAP --> 非self_visit --> Facebook --> 其他 --> 贡献率为 -1.99; 非WAP --> s20220303	duv		
-0.5174547421940111	其他 : 1.275; WAP: 1.063; 付费SNS: 0.597; PPC: 0.208; self_visit: 0.031; Fac	非WAP --> 非self_visit --> Facebook --> 其他 --> 贡献率为 -1.92; 非WAP --> s20220304	duv		
-0.5262356875372315	其他 : 1.184; WAP: 0.961; 付费SNS: 0.666; PPC: 0.204; self_visit: 0.030; Fac	非WAP --> 非self_visit --> Facebook --> 其他 --> 贡献率为 -1.89; 非WAP --> s20220305	duv		
-0.5265053087693369	其他 : 1.391; WAP: 1.028; 付费SNS: 0.598; PPC: 0.223; free: 0.033; self_vis	非WAP --> 非self_visit --> Facebook --> 其他 --> 贡献率为 -1.86; 非WAP --> s20220306	duv		
-0.5211738174631667	其他 : 1.355; WAP: 0.968; 付费SNS: 0.655; PPC: 0.240; self_visit: 0.026; Fac	非WAP --> 非self_visit --> Facebook --> 其他 --> 贡献率为 -1.85; 非WAP --> s20220307	duv		
-0.5140445976873306	其他 : 1.423; WAP: 1.024; 付费SNS: 0.613; PPC: 0.215; free: 0.035; self_vis	非WAP --> 非self_visit --> 非Facebook --> SNS_rmk --> 其他 --> 贡献率为 -1.120220308	duv		
-0.5083027541981548	其他 : 1.425; WAP: 1.023; 付费SNS: 0.613; PPC: 0.217; free: 0.036; self_vis	非WAP --> 非self_visit --> 非Facebook --> SNS_rmk --> 其他 --> 贡献率为 -1.120220309	duv		
-0.5029945712735063	其他 : 1.351; WAP: 0.960; 付费SNS: 0.669; PPC: 0.234; self_visit: 0.029; Fac	非WAP --> 非self_visit --> Facebook --> 其他 --> 贡献率为 -1.79; 非WAP --> s20220310	duv		
-0.5054790406232971	其他 : 1.350; WAP: 0.959; 付费SNS: 0.669; PPC: 0.236; self_visit: 0.030; Fac	非WAP --> 非self_visit --> Facebook --> 其他 --> 贡献率为 -1.78; 非WAP --> s20220311	duv		
-0.5051243155128544	其他 : 1.350; WAP: 0.957; 付费SNS: 0.665; PPC: 0.240; self_visit: 0.032; Fac	非WAP --> 非self_visit --> Facebook --> 其他 --> 贡献率为 -1.77; 非WAP --> s20220312	duv		
-0.5024386604646413	其他 : 1.342; WAP: 0.955; 付费SNS: 0.661; PPC: 0.241; self_visit: 0.033; Fac	非WAP --> 非self_visit --> Facebook --> 其他 --> 贡献率为 -1.77; 非WAP --> s20220313	duv		
-0.4984156550597358	其他 : 1.336; WAP: 0.953; 付费SNS: 0.654; PPC: 0.245; self_visit: 0.034; Fac	非WAP --> 非self_visit --> Facebook --> 其他 --> 贡献率为 -1.76; 非WAP --> s20220314	duv		
-0.4949180504598089	WAP: 1.525; 其他 : 1.365; pay: 0.873; 付费SNS: 0.152; TAC广告外投: 0.081; 非WAP	非WAP --> 非self_visit --> Facebook --> 其他 --> 贡献率为 -1.73; 非WAP --> s20220315	duv		
-0.4851857581731981	WAP: 1.521; 其他 : 1.360; pay: 0.874; 付费SNS: 0.154; TAC广告外投: 0.080; 非WAP	非WAP --> 非self_visit --> Facebook --> 其他 --> 贡献率为 -1.70; 非WAP --> s20220316	duv		
-0.4772591112265196	其他 : 1.240; WAP: 1.157; pay: 0.608; 付费SNS: 0.247; self_visit: 0.038; dir	非WAP --> 非self_visit --> Facebook --> 其他 --> 贡献率为 -1.67; 非WAP --> s20220317	duv		
-0.4730764080781817	WAP: 1.222; 其他 : 1.158; pay: 0.663; 付费SNS: 0.233; self_visit: 0.042; dir	非WAP --> 非self_visit --> Facebook --> 其他 --> 贡献率为 -1.66; 非WAP --> s20220318	duv		
-0.4703093526891138	WAP: 1.220; 其他 : 1.154; pay: 0.665; 付费SNS: 0.233; self_visit: 0.042; dir	非WAP --> 非self_visit --> Facebook --> 其他 --> 贡献率为 -1.65; 非WAP --> s20220319	duv		

图十一 每日duv异动归因结果表

更多2022财年团队数据官暨FBI达人赛优秀案例：<https://ata.alibaba-inc.com/articles/235703>

附录1

证明 加法贡献算法

已知 $Y = \sum_i X_i$

贡献等于 $C_{X_i} = \frac{X_i^1 - X_i^0}{Y^0} = \frac{\Delta X_i}{Y^0}$,

$\Delta Y\% = \frac{Y^1 - Y^0}{Y^0} = \frac{\sum_i X_i^1 - \sum_i X_i^0}{Y^0} = \frac{\sum_i X_i^1 - X_i^0}{Y^0} = \frac{\sum_i \Delta X_i}{Y^0} = \sum C_{X_i}$

证明 乘法贡献算法

已知 $Y = \prod_i X_i$, $L(Y^1, Y^0) = \frac{Y^1 - Y^0}{\ln(Y^1) - \ln(Y^0)} = \frac{\Delta Y}{\ln(Y^1) - \ln(Y^0)}$

贡献等于 $C_{X_i} = \frac{L(Y^1, Y^0) \cdot \ln(\frac{X_i^1}{X_i^0})}{Y^0}$,

目标波动 $\Delta Y\% = \frac{Y^1 - Y^0}{Y^0} = \frac{\Delta Y \cdot \ln(\frac{Y^1}{Y^0})}{Y^0 \cdot \ln(\frac{Y^1}{Y^0})} = \frac{\Delta Y \ln(\frac{\prod X_i^1}{\prod X_i^0})}{Y^0 \cdot \ln(\frac{Y^1}{Y^0})} = \frac{\Delta Y \sum_i \ln(\frac{X_i^1}{X_i^0})}{Y^0 \cdot \ln(\frac{Y^1}{Y^0})} = \sum_i \frac{L(Y^1, Y^0)}{Y^0} \ln(\frac{X_i^1}{X_i^0}) = \sum_i C_{X_i}$

该方法为乘积因子拆解 (Logarithmic Mean Index Method), 符合MECE原则, 对目标指标变化不重不漏地拆解到各个链路, 贡献率总和为乘积结果的变化率

证明见: Ang, Beng W., F. Q. Zhang, and Ki-Hong Choi. "Factorizing changes in energy and environmental indicators through decomposition." Energy 23.6 (1998): 489–495.

证明 比率贡献算法

$$\text{已知 } Y = \frac{S}{P} = \frac{\sum_i s_{X_i}}{\sum_i p_{X_i}}, \quad P_i = \frac{p_{X_i}}{\sum_i p_{X_i}}, \quad S_i = \frac{s_{X_i}}{\sum_i s_{X_i}}, \quad Y_i = \frac{s_{X_i}}{p_{X_i}}$$

$A_{X_i} = (Y_i^1 - Y_i^0) \times P_i^0$, 其中 Y_i^1 是当月的数据, Y_i^0 为上个月 (同比/环比) 的数据。

$$B_{X_i} = (P_i^1 - P_i^0) \times (Y_i^1 - Y^0)$$

$$\text{贡献等于 } C_{X_i} = \frac{A_{X_i} + B_{X_i}}{Y^0}$$

算法得到的贡献率依旧相互独立, 符合MECE原则, 且通过分别观测 A_{X_i} 和 B_{X_i} 有助于我们避开辛普森悖论带来的陷阱。

$$\begin{aligned} \Delta Y\% &= \frac{Y^1 - Y^0}{Y^0} = \frac{\frac{\sum_i s_{X_i}^1}{\sum_i p_{X_i}^1} - \frac{\sum_i s_{X_i}^0}{\sum_i p_{X_i}^0}}{Y^0} = \frac{\frac{\sum_i \frac{s_{X_i}^1}{p_{X_i}^1} \cdot p_{X_i}^1}{\sum_i p_{X_i}^1} - \frac{\sum_i \frac{s_{X_i}^0}{p_{X_i}^0} \cdot p_{X_i}^0}{\sum_i p_{X_i}^0}}{Y^0} \\ &= \frac{\sum_i \frac{s_{X_i}^1}{p_{X_i}^1} \cdot \frac{p_{X_i}^1}{\sum_i p_{X_i}^1} - \sum_i \frac{s_{X_i}^0}{p_{X_i}^0} \cdot \frac{p_{X_i}^0}{\sum_i p_{X_i}^0}}{Y^0} = \frac{\sum (P_i^1 \cdot Y_i^1) - \sum (P_i^0 \cdot Y_i^0)}{Y^0} = \frac{\sum (P_i^1 \cdot Y_i^1) - \sum (P_i^1 \cdot Y^0) - \sum (P_i^0 \cdot Y_i^0) + \sum (P_i^0 \cdot Y^0)}{Y^0} \\ &= \frac{(\sum (P_i^0 \cdot Y_i^1) - \sum (Y_i^0 \cdot P_i^0)) + (\sum (P_i^1 \cdot Y_i^1) - \sum (P_i^1 \cdot Y^0) - \sum (P_i^0 \cdot Y_i^1) + \sum (P_i^0 \cdot Y^0))}{Y^0} = \frac{\sum A_{X_i} + B_{X_i}}{Y^0} = \sum C_{X_i} \end{aligned}$$

备注: C_{X_i} 符合mece原则证明: 相互独立: A_{X_i} 和 B_{X_i} 的计算公式中不涉及其他分项的完全穷尽。

参考

1. Ang, Beng W., F. Q. Zhang, and Ki-Hong Choi. "Factorizing changes in energy and environmental indicators through decomposition." Energy 23.6 (1998): 489–495.
2. Ang B W . The LMDI approach to decomposition analysis: a practical guide[J]. Energy Policy, 2005, 33(7):867–871.
3. [《波动解读—指标拆解的加减乘除双因素》](#)
4. [《决策树 \(decision tree\) \(二\)——剪枝》](#)
5. [《Minimal Cost–Complexity Pruning》](#)