AnyWhy 用户使用手册

北京格瑞纳电子产品有限公司

目录

1 AnyWhy 软件布局及功能 **2**

1.1 启动 2

1.2 主页 4

1.3 项目管理 5

1.4 项目编辑及运行 6

1.5 节点 10

1.5.1 流程开始节点 10

1.5.2 流程结束节点 10

1.5.3 判断节点 10

1.5.4 因果模型构建 11

1.5.5 路径分析 13

1.5.6 路径判定 14

1.5.7 平均因果效应 16

1.5.8 直接因果效应 18

1.5.9 内在因果效应 21

1.5.10 分布变化归因 23

1.5.11 干预分析 26

1.5.12 反事实分析 29

1.5.13 独立性分析 31

1.5.14 异常归因 32

1.5.15 异常数据选择 34

2 功能模块 **37**

2.1 模型分析（Model Analysis） 37

2.1.1 识别前门路径（frontdoor path） 37

2.1.2 识别后门路径（backdoor path） 37

2.2 独立性分析（Independence Analysis） 38

2.3 直接因果效应（Arrow Strength） 38

2.4 内在因果效应（Intrinsic Causal Influence） 39

2.5 平均因果效应（Average Treatment Effect or Average causal Effect） 39

2.6 干预分析（intervention analysis） 39

2.7 分布变化归因（attributing distributional change） 40

2.8 反事实分析（counterfactual analysis） 40

2.9 异常归因（anomaly attribution） 41

2.10 因果发现（causal learn） 41

2.10.1 基于约束的因果发现方法 41

2.10.1.1 Perform Peter-Clark（PC） 41
2.10.1.2 Fast Causal Inference (FCI) 42
2.10.2 基于分数的因果发现方法 (Score-based causal discovery methods) 43
2.10.2.1 GES with the BIC score or generalized score . . . 43
2.10.2.2 Exact Search： 44

3 案例 45
3.1 案例背景 45
3.2 创建项目 45
3.3 导入数据、模型建模 47
3.3.1 手动建立模型 47
3.3.2 使用因果发现生成因果模型 54
3.4 独立性分析 56
3.5 路径分析 58
3.6 直接因果效应 60
3.7 内在因果效应 63
3.8 平均因果效应 64
3.9 干预分析 66
3.10 反事实分析 67
3.11 异常归因 70
3.12 分布变化归因 73

摘要

近年来，因果推断在学术界和工业界都引起了广泛的关注。在学术界中，因果推断是非常热门的研究方向之一，相关学术研究成果近两年占比明显增加。在工业界，因果推断相关技术在推荐系统、用户增长等领域也得到了广泛的应用。

为了使因果推断能够使在多领域实现普及，AnyWhy 内置因果推断工具，用户只需要填入所需参数，即可实现因果推断相关工作。

AnyWhy 的商业优势也很明显。它是：

1. 基于微软与亚马逊联合开发的 dowhy 库
2. 可作为各领域决策支持工具
3. 可以考虑多种决策方案
4. 量化各因素之间的因果关系

1 AnyWhy 软件布局及功能

1.1 启动

在 AnyWhy 文件夹下有以下几个文件：

名称	修改日期	类型	大小
app_run	2025/5/7 17:23	文件夹	
client	2025/5/7 15:32	文件夹	
end.bat	2024/12/26 10:43	Windows 批处理...	1 KB
start.bat	2025/5/7 16:20	Windows 批处理...	1 KB

图 1: AnyWhy 文件夹

双击 start.bat 脚本即可打开服务

图 2: start.bat

服务成功启动将会出现两个窗口，一个服务正常运行窗口，一个 backend 的黑窗口。窗口表现如下：

图 3: 服务正在运行

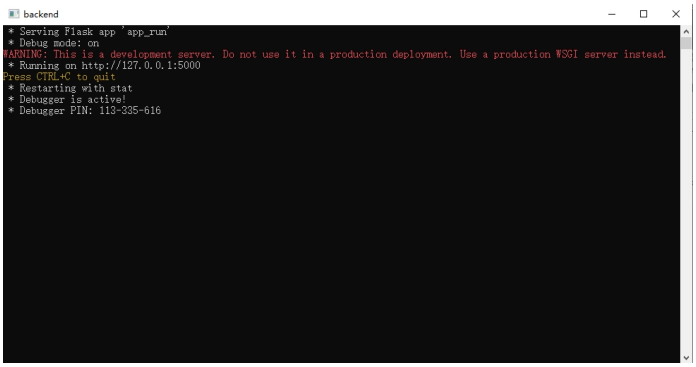


图 4: backend 黑窗口

选择任意浏览器，在域名窗口键入 **localhost:8090**

图 5: 域名窗口

成功进入主页

图 6: 域名加主页

点击**开始使用**跳转到登录界面，**默认账户名称为 admin，密码为 admin**

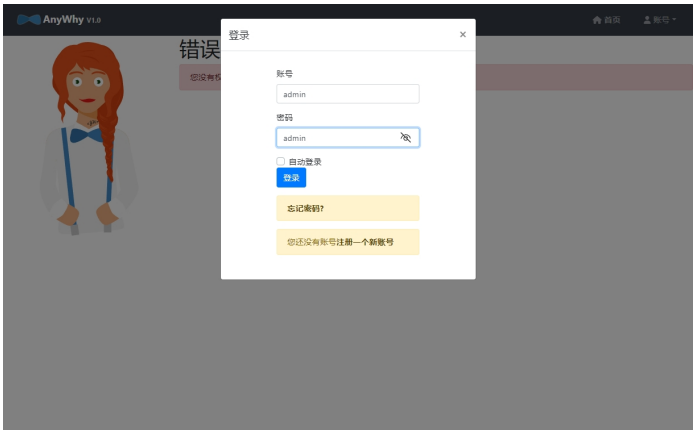


图 7: 主页

登录成功后，主页界面右上角将出现**项目**以及**管理**按钮。

图 8: 登陆后主页

1.2 主页

图 9: 主页

主页可以选择查看功能说明。

主要功能为路径分析（path analysis）、平均因果效应（ATE or Average Treatment Effect）、直接因果效应（arrow strength）、内在因果效应（Intrinsic Causal Influence）、反事实分析（counterfactual analysis）、干预分析（intervention analysis）、分布变化归因（attributing distributional change）、异常归因（anomaly attribution）、独立性分析（Independence analysis）点击路径分析（path analysis），可以进入路径分析（path analysis）说明界面，查看路径分析相关内容。

点击开始使用或者右上角登陆即可登录账户，**目前账户用户名为 admin，密码为 admin**。

登陆后主页右上角出现项目以及管理按钮。

图 10: 登陆后主页

1.3 项目管理

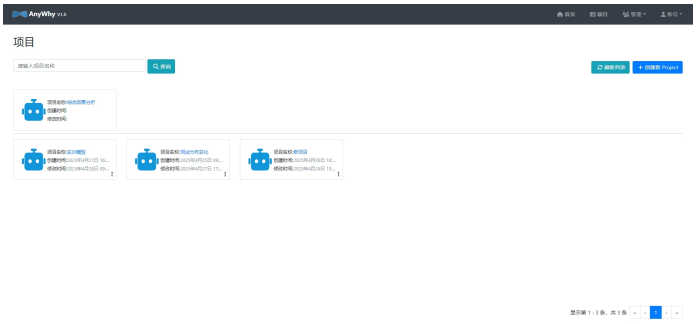


图 11: 项目管理

进入项目管理界面后，可以看到已创建的项目，第一行为示例项目，名称为综合因果分析。使用右上角的创建新 project 即可创建一个新的项目。

1.4 项目编辑及运行

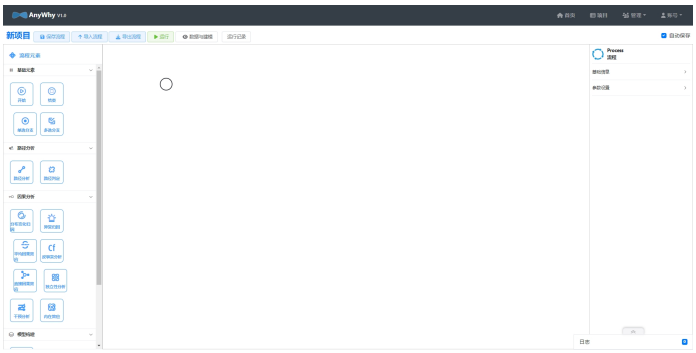


图 12: 项目编辑模式

新建项目完成后，点击项目进入项目编辑界面。位于上方的“保存流程”按钮可以保存已建好的流程，“导出流程”按钮可以将已建立好的流程导出为.xml 格式文件，“导入流程”按钮可以导入已保存好的.xml 流程文件，“运行”按钮可将项目由编辑模式转变为运行模式。

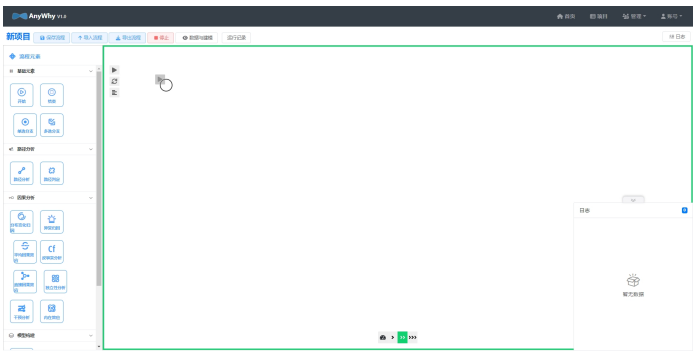
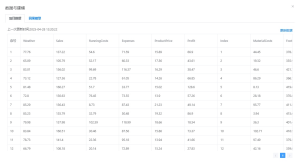


图 13: 项目运行模式

“导入数据”按钮可以从本地将数据进行导入，可支持格式为 excel 以及 csv，导入完成后“导入数据”按钮将变为“查看数据”按钮，点击即可查看上传的数据。点击“更新数据”可以选择新数据。

图 14: 查看数据



编号	Weather	Sales	StockingLevel	Expenses	ProductPrice	Profit	Index	MaterialLevels	Fact
1	75.76	115.12	54.6	71.91	19.84	49.6	1	44.45	19.[illegible]
2	43.44	126.75	12.17	86.11	17.66	43.61	2	54.31	3.0[illegible]
3	82.07	126.12	38.48	7.31.17	34.29	33.87	3	48.6	42.7[illegible]
4	75.72	77.39	22.95	19.55	34.24	66.05	4	39.25	28[illegible]
5	41.46	146.11	14.1	33.77	19.61	5.84	5	9.11	49.[illegible]
6	7.14	141.61	74.42	71.52	1.11	17.41	6	28.16	15.[illegible]
7	45.29	138.41	8.71	37.41	2.12	49.14	7	35.77	47.[illegible]
8	43.21	11.79	11.75	35.48	31.12	34.3	8	3.84	47.[illegible]
9	75.92	117.38	72.23	7.52.03	19.84	72.24	9	39.1	47.[illegible]
10	41.66	146.11	35.48	47.34	19.84	71.27	10	14.17	49.[illegible]
11	76.75	14.14	21.34	35.14	1.14	4.140	11	67.45	15.[illegible]
12	44.79	138.31	20.14	72.30	71.14	2.52	12	42.16	3.0[illegible]

图 15: 项目数据

在项目编辑模式界面中，左侧栏为节点区域，使用鼠标选择所需节点后，将节点拖入中间的操作区，即可完成对节点使用。

右侧栏为属性设置区域（一般不会用到）。上方的“运行记录”按钮可以查看之前每一次运行的结果。

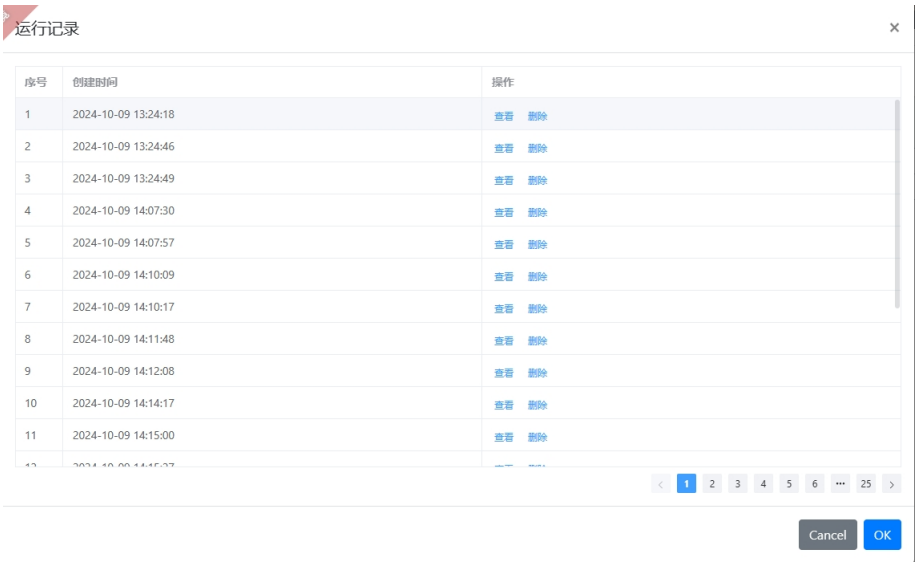


图 16: 运行记录

当勾选上位于“运行记录”旁边的“自动保存”按钮时，只要工作流有修改，便会自动保存。

图 17: 自动保存

在项目运行的界面中，左侧的节点区消失。在上方的运行记录旁出现日志按钮，点击日志按钮会在屏幕右侧出现“运行日志”界面，在此界面可以查看各节点的结果，点击“生成报告”可以将节点结果导出。

图 18: 日志

图 19: 运行日志

1.5 节点

1.5.1 流程开始节点

管控流程开始的节点，所有的流程都需要以这个节点开始。

图 20: 流程开始节点

当处于项目运行的界面中，流程开始节点会出现运行按钮，点击运行按钮可以开始一个流程。

图 21: 流程开始节点-运行模式

1.5.2 流程结束节点

管控流程结束的节点，所有的流程都需要以这个节点结束。

图 22: 流程结束节点

当处于项目运行的界面中，流程会流向流程结束节点，完成后在流程结束节点上方出现 finished。

图 23: 流程结束节点-运行模式

1.5.3 判断节点

控制流程的走向，起到选择功能的作用。

图 24: 判断节点

当处于项目运行界面中，流程到达判断节点时，会出现判断选择弹窗，弹窗选项为后续路径名称。

图 25: 判断选择弹窗

1.5.4 因果模型构建

在项目运行界面中，流程到达因果模型构建节点时，会弹出创建模型界面，界面会自动读取录入数据的表头，将其转换为节点，用户可以使用这些节点构建因果图。每个节点可以选择它与其父节点的链接关系，但**当节点为根节点时，及没有任何节点指向它时，该节点设置的链接关系不会影响计算结果**。

图 26: 因果模型构建节点

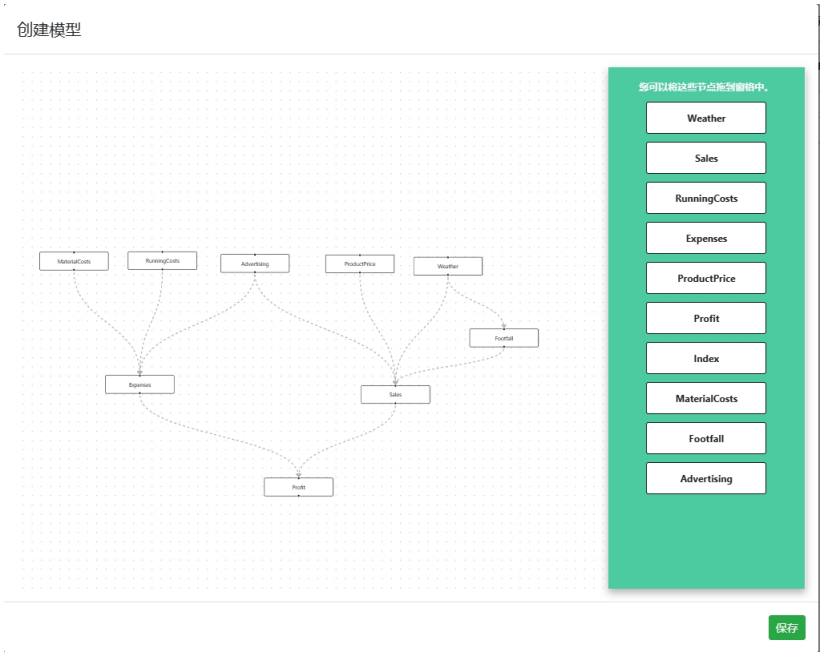

创建模型
MaterialCosts
RunningCosts
Advertising
ProductPrice
Weather
Footfall
Expenses
Sales
Profit
多可以将这些节点拖到窗格中。
Weather
Sales
RunningCosts
Expenses
ProductPrice
Profit
Index
MaterialCosts
Footfall
Advertising
保存

图 27: 创建模型

以下节点，可以在右侧边栏设置参数，也可以在流程运行的时候在弹窗中进行参数设置。

1.5.5 路径分析

图 28: 路径分析节点

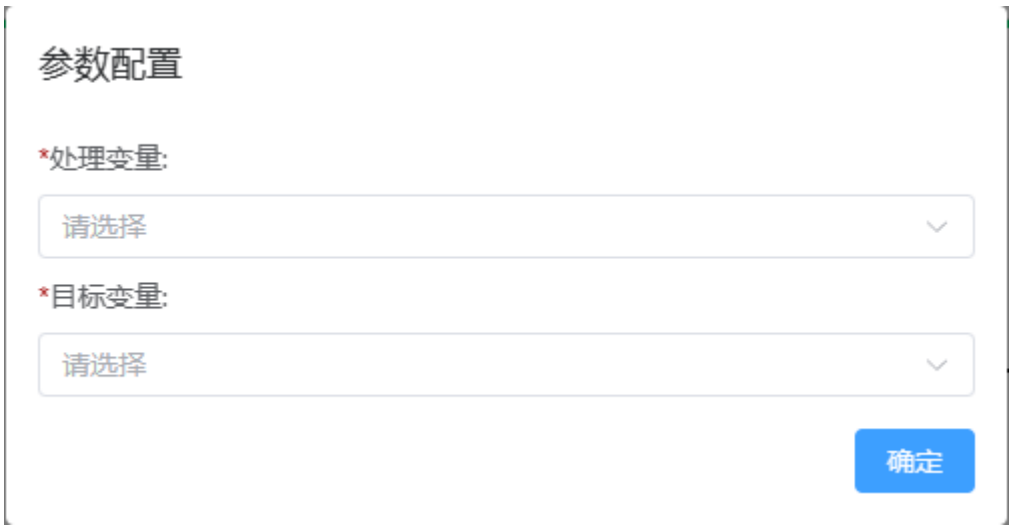


图 29: 路径分析-参数配置（弹窗）

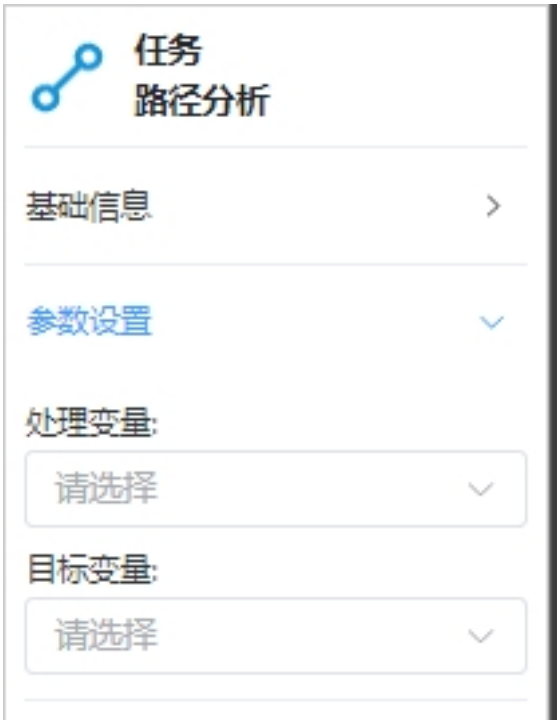


图 30: 路径分析-参数配置（边栏）

处理变量：需要发生变化的变量，例如：运营成本等
目标变量：关注因素，例如：公司利润等

1.5.6 路径判定

图 31: 路径判定

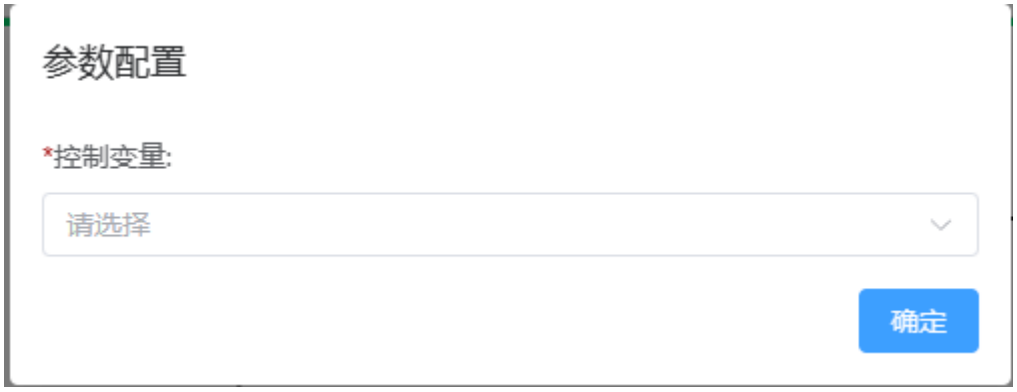


图 32: 路径判定-参数配置

控制变量：需要控制的因素，所有经过该因素的路径将会被阻断。

1.5.7 平均因果效应

图 33: 平均因果效应节点

图 34: 平均因果效应-参数配置

图 35: 平均因果效应-参数配置

处理变量：需要发生变化的变量，例如：运营成本等
原始值（Reference）：变化前的处理变量数值
替代值（Alternative）：变化后的处理变量数值

目标变量：关注因素，例如：公司利润等

高级参数：

图 36: 平均因果效应-高级参数

观测数据：使用录入的数据来计算 ate

抽样样本数量（默认计算方式为抽样样本数量）：使用拟合好因果模型中的样本进行计算，规定使用的样本数量

1.5.8 直接因果效应

图 37: 直接因果效应节点

图 38: 直接因果效应-参数配置

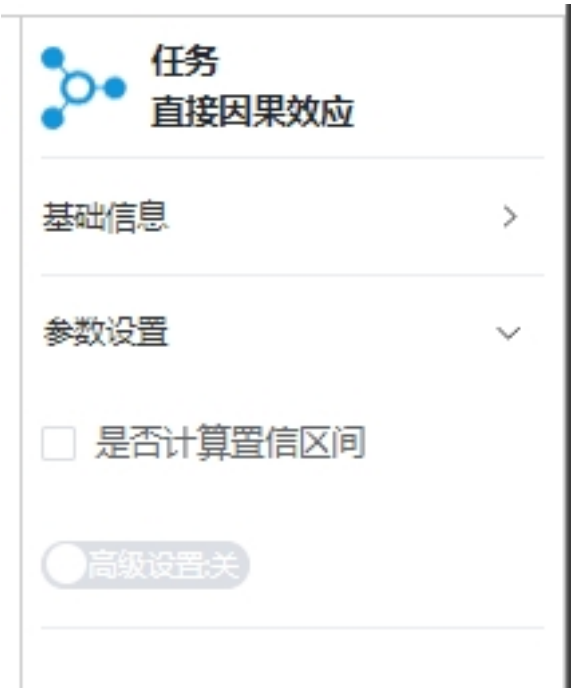


图 39: 直接因果效应-参数配置（边栏）

是否计算置信区间：勾选将从因果模型中抽取多个样本（默认为 20 个），随后从抽取样本中得到直接因果效应置信区间；不勾选将计算多次后取平均值（默认为 3 次）

高级参数：

图 40: 直接因果效应-高级参数

样本数量：使用拟合好因果模型中的样本进行计算，规定使用的样本数量

置信水平：用于评估置信区间

抽样次数：设定从因果模型中抽取的样本数量

最大运行次数：运算直接因果效应的次数，计算平均值（默认为 3 次）

最大公差：如果连续两次运行之间估计强度的百分比变化低于指定公差，则算法将在达到最大运行数之前终止。值为 0.01 表示变化小于 1%。然而，为了最小化随机性的影响，必须至少连续运行三次，其中变化低于阈值。

评估方法：计算直接因果效应方法（目标变量为连续变量时，默认为方差之差；目标变量为分类变量时，默认为 KL 散度），应填入 python 函数形式代码（应填入两个变量的函数），例如：

```
def add(a,b):
    return a + b
```

1.5.9 内在因果效应

图 41: 内在因果效应节点

图 42: 内在因果效应-参数配置

是否计算置信区间：勾选将从因果模型中抽取多个样本（默认为 20 个），随后从抽取样本中得到内在因果效应置信区间

高级参数：

图 43: 内在因果效应-高级参数

置信水平：用于评估置信区间
抽样次数：设定从因果模型中抽取的样本数量
运行次数：运算内在因果效应的次数，计算平均值（默认为 3 次）
预测模式：

Approx：计算近似内在因果效应结果（消耗时间较长）

Exact：计算精确内在因果效应结果（消耗时间较短）

shapley 评估方法：评估 shapley 指标（默认为方差），填入形式应为 python 形式函数代码（应填入两个变量的函数）例如：

```
def add(a,_):
    return mean(a)
```

1.5.10 分布变化归因

图 44: 分布变化归因节点

图 45: 分布变化归因-参数配置

图 46: 分布变化归因-参数配置（边栏）

目标变量：关注因素，例如：公司利润等
数据分割索引：选择数据分割的节点，将录入的数据集分为两份，随后比较这两份数据集的分布变化。
是否计算置信区间：勾选将从因果模型中抽取多个样本（默认为 20 个），随后从抽取样本中得到内在因果效应置信区间。

高级参数:

图 47: 分布变化归因-高级参数

估计 shapely 值样本数: 用于估计 shapely 值时使用的样本数量，默认为 1000
置信水平: 用于评估置信区间
抽样次数: 计算置信区间时的次数，运算出结果最大值为上限，最小值为下限（默认为 20 次）
评估方法: 评估 shapley 指标（默认为方差），填入形式应为 python 形式函数代码（应填入两个变量的函数）例如:

```
def add(a,_):
    return mean(a)
```

1.5.11 干预分析

图 48: 干预分析节点

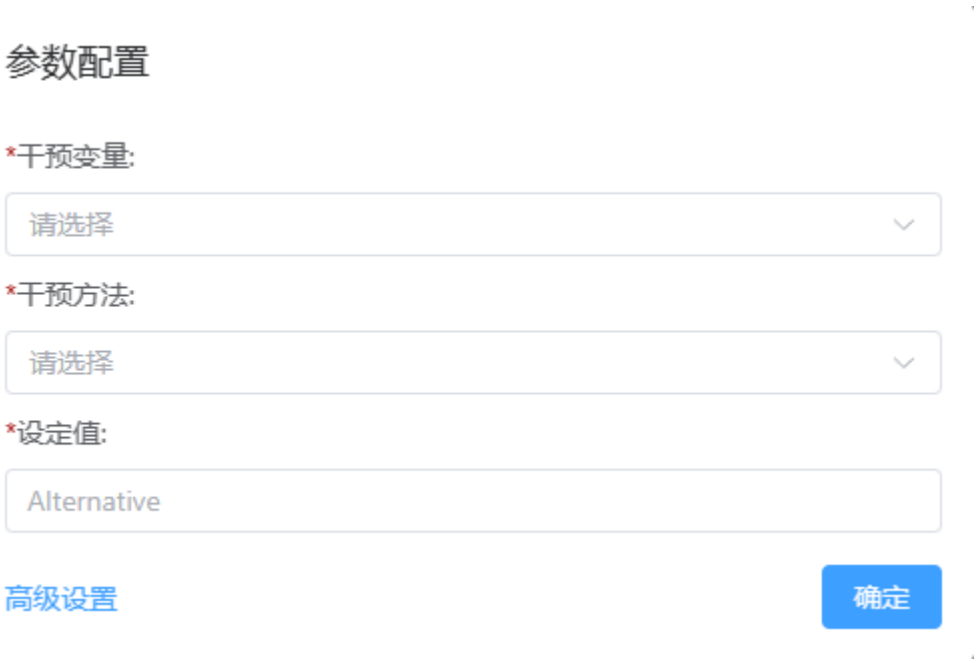


图 49: 干预分析-参数配置

图 50: 干预分析-参数配置

干预变量：需要改变的变量名称

干预方法：

1. **原子干预**：将干预变量设定为一个固定值，使群体中每个个体的干预变量都相同。这种方法可以消除干预变量对目标变量的影响。

2. **移位干预**：将干预变量增加一个设定的值，使群体中每个个体的干预因素都增加相同的值。这种方法可以用来观察干预变量对目标变量的影响。

设定值：设定干预变量的数值

高级设置:

图 51: 干预分析-高级参数

观测数据: 使用录入的数据来计算干预分析值
抽样样本数量 (默认计算方式为抽样样本数量): 使用拟合好因果模型中的样本进行计算，规定使用的样本数量

1.5.12 反事实分析

图 52: 反事实分析节点

图 53: 反事实分析-参数配置

图 54: 反事实分析-参数配置

干预变量：需要改变的变量名称。
设定值：将干预变量的值修改为设定值。
数据索引：想要更改的数据对应索引值。

1.5.13 独立性分析

图 55: 独立性分析节点

高级设置:

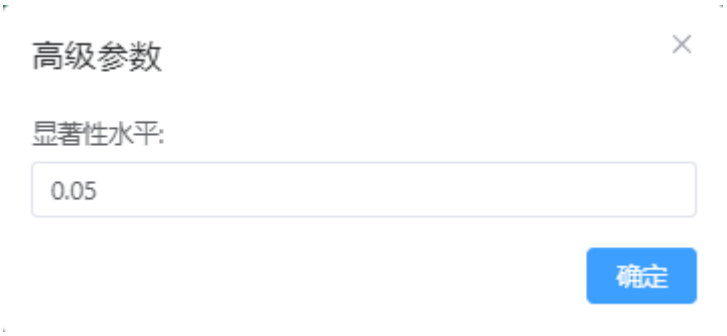


图 56: 独立性分析-高级设置

显著性水平: 用于评估置信区间

1.5.14 异常归因

图 57: 异常归因节点

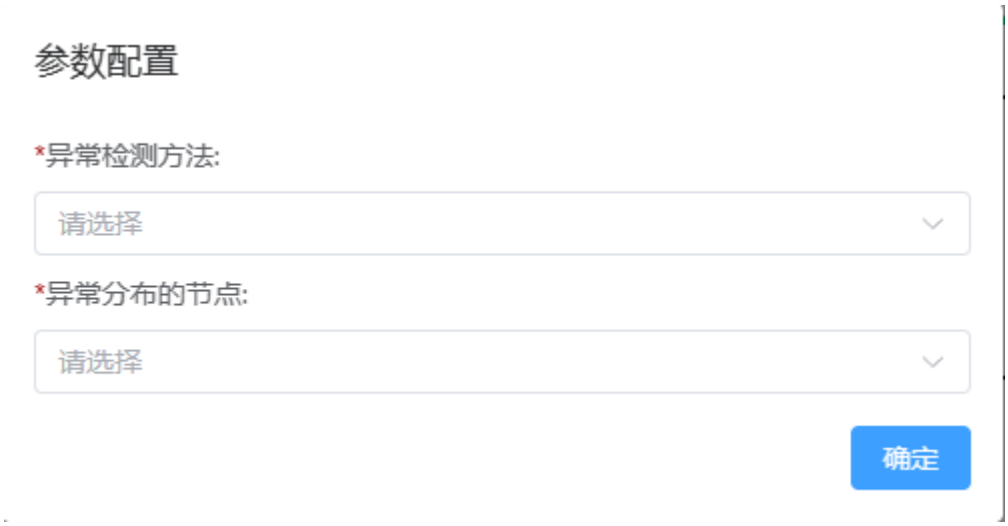


图 58: 异常归因-参数配置

图 59: 异常归因-参数配置

异常检测方法:

1. **ECOD**: 基于经验累积分布的异常检测。首先以非参数方式估计变量的

分布。然后，它将所有维度的估计尾部概率相乘，以获得观测值的异常得分。

2. **KNN**：K-最近邻算法是一种基本的分类与回归方法，同时也可以用来检测异常值。

3. **IForest**：孤立森林算法，其原理基于异常点密度小，基于树模型容易被一下切割出来，正常值密度大，需要切割多次才能得到目标值。

异常分布的节点：目标节点，选择需要检测异常的因素。

1.5.15 异常数据选择

异常归因参数配置计算完成后弹出）

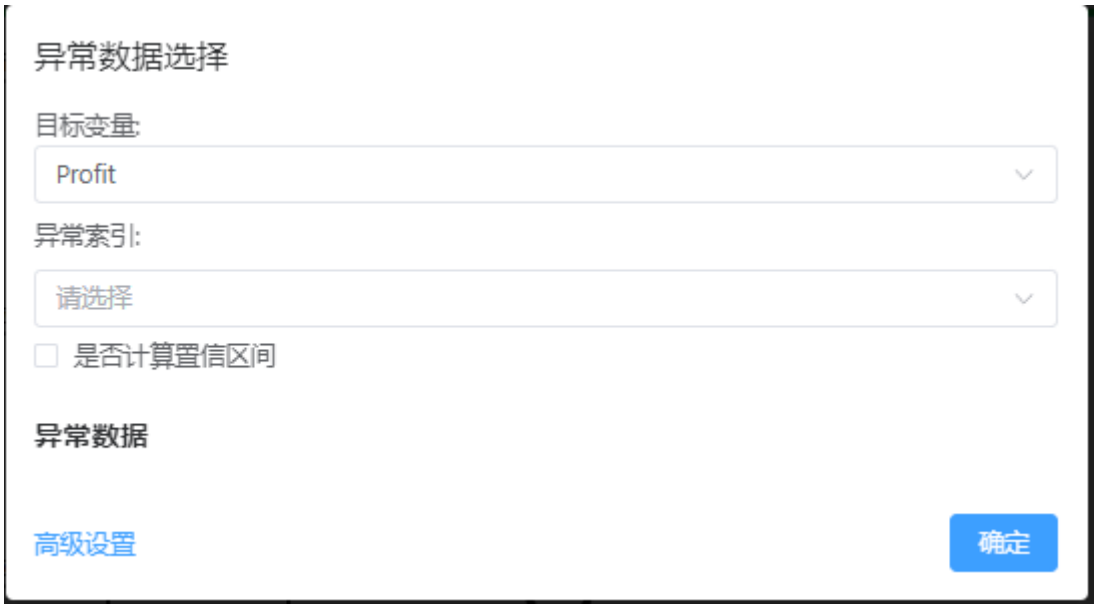


图 60: 异常数据选择-参数配置

目标变量：目标节点，选择需要检测异常的因素。

异常索引：从检测出的异常数据集中选择需要计算的数据索引。

高级参数:

图 61: 异常数据选择-高级参数

样本数量：从因果模型中提取样本数，用于计算各因素的边缘分布，完成对异常分数的分配。

置信水平：用于评估置信区间。

异常分数计算方法（默认为 MedianCDFQuantileScorer）：

MedianCDFQuantileScorer：

给定异常观测值 x 和 x 分布的样本，该分数表示：

score(x) = 1 - 2 * min[P(X > x) + P(X = x) / 2, P(X < x) + P(X = x)/2]

这个表达式意思是样本 x 的值是分布 X 中的一部分，当 x 靠近 X 分布的两侧，即越远离 X 分布的中位数时，score(x) 越大

RescaledMedianCDFQuantileScorer：

给定异常观测值 x 和 x 分布的样本，该分数表示：

score(x) = -log(2 * min[P(X > x) + P(X = x) / 2, P(X < x) + P(X=x)/2])

MeanDeviationScorer：

给定异常观测值 x 和 X 分布的样本，该分数表示：

score(x) = |x - E[X]| / std[X]

该表达式表示为给定样本 x 到 X 分布均值的距离并按 X 分布的标准差进行缩放，这也相当于高斯变量中的 z 分数。

分数越高，观测值与 X 分布均值的偏差越大。

MedianDeviationScorer：

给定异常观测值 x 和 X 分布的样本，该分数表示：

score(x) = |x - med[X]| / mad[X]

该表达式表示为给定样本 x 到 X 分布中位数的距离对其进行评分，并按 X 分布的中位数绝对偏差进行缩放。

分数越高，观测值与 X 分布中位数的偏差越大。

InverseDensityScorer：

使用 1/p(x) 作为异常得分，其中 x 为需要评分的数据。默认使用高斯混合模型作为密度估计模型。

2 功能模块

本模块将介绍因果 AI 平台的所有功能。主要功能包括：模型分析（Model Analysis），直接因果效应（arrow strength），内在因果效应（Intrinsic Causal Influence），平均因果效应（ATE or ACE），干预分析（intervention analysis），分布变化归因（attributing distributional change），反事实分析（counterfactual analysis），异常归因（anomaly attribution），因果发现（causal learn）

2.1 模型分析（Model Analysis）

2.1.1 识别前门路径（frontdoor path）

在因果图中，如果变量集 Z 满足以下条件，则称 Z 满足（X，Y）的前门准则，X 到 Y 的路径为前门路径：

1. 阻断了 X 到 Y 的所有路径（即 Z 切断了 X 对 Y 的直接影响。）

2.X 到 Z 之间没有未阻断的路径

3.Z 到 Y 之间的所有后门路径都被 X 阻断

图 62: 前门路径

2.1.2 识别后门路径（backdoor path）

在因果图中，从目标变量 Y 到干预变量 X 之间存在的不以 X 为起点的路径，且这些路径中包含指向 X 的箭头。

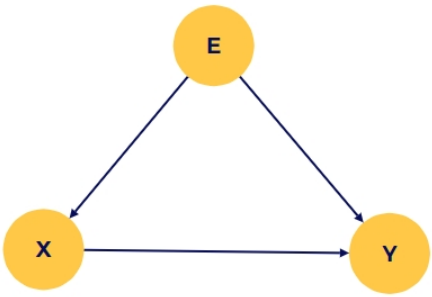


图 63: 后门路径

由于后门路径的存在可能导致因果关系的误判，因为混杂因素（E）可能同时影响 X 和 Y。例如，为了验证某种药物是否能使病人痊愈，我们需要剔除病人性别对药物效果的影响。因此，我们需要使用策略来阻断后门路径。

控制变量：通过控制可能同时影响 X 和 Y 的变量 E（即混杂因素），可以阻断后门路径。

do 算子：在某些情况下，可以通过干预手段，使得 X=x 时 Y=y 的概率，反映了如果群体中的每个个体均将 X 值固定为 x 时，Y 的总体分布。

使用工具变量：在某些情况下，可以使用与 X 相关但与 Y 无直接因果关系的工具变量 S 来估计 X 对 Y 的因果效应，从而避开后门路径的影响。

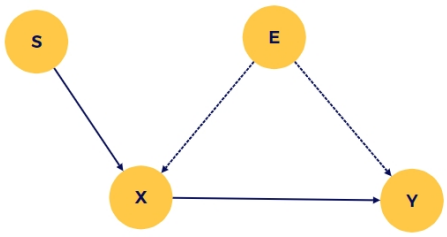


图 64: 后门路径

2.2 独立性分析（Independence Analysis）

此方法旨在验证因果图的准确性。

独立性检验：使用基于核的条件独立性检验，判断两个变量是否独立，从而了解两个变量节点是否应该相连。当计算出的 p 值或者调整 p 值小于 0.05 时，两个变量不独立，两个变量节点之间应该相连。

局部马尔可夫检验：通过判断变量节点是否满足局部马尔科夫性，从而判断其他未与它相连的节点是否应该与其相连。当计算出的节点 p 值或者调整 p 值为大于 0.05 时，节点与其非相邻节点独立，即不应该相连。

2.3 直接因果效应（Arrow Strength）

此方法是一种量化因果效应的方式。通过移除某个影响因素与目标变量之间的连接关系，观察移除前后分布的变化，然后使用某种度量来评估分布变化，从而量化因果效应。在默认情况下，我们使用去除连接关系前后方差的差异或相对熵作为度量标准。例如：影响利润的因素有销售额和总成本。一、首先根据录入的数据设置好因果效应链接关系生成因果模型。二、根据已有数据和因果模型生成各因素数据分布图。三、固定销售额的值观察总成本正常变化时总利润分布的变化情况计算两个利润分布的方差之差就为总成本对总利润的因果效应强度四、随后重复上述操作固定总成本的分布观察总利润分布的变化情况

2.4 内在因果效应（Intrinsic Causal Influence）

此方法用于在剔除祖先节点的影响后，量化影响因素对目标变量的因果效应。在某些情况下，影响目标变量的往往不是直接的影响因素，而是影响因素的祖先节点。一些上游影响因素只是连接祖先节点和目标变量的桥梁，对目标变量的直接影响较小。为了区分这种情况，内在因果效应能够将影响因素本身的因果效应与从其祖先获得的效应分离开来，然后通过方差或熵等度量方式来量化，从而确定影响因素本身对目标变量的因果效应。一：首先，同直接因果效应相同，根据录入的数据和链接方式，我们生成因果模型。例如：总利润与销售额和总成本模型。二：根据已有数据分布和因果模型，生成各节点的噪声。例如：将总利润的数据减去总成本再加上销售额，剩下的数据即为总利润的噪声。注意：我们默认根节点的噪声为数据本身。三：根据链接方式以及因果模型，求得各节点噪声，四：随后根据获得的噪声和链接方式，得到各因素的预测分布。这里，我们在获得销售额和总成本的预测分布后，根据拟合好的因果模型方程，结合之前获得的利润噪声分布，可以获得预测利润分布。五：使用 shapley 值法，根据生成的预测分布以及链接方式，算出每个因素为总利润方差的贡献度。其中，shapley 值法是机器学习常用的贡献度计算方法。

2.5 平均因果效应（Average Treatment Effect or Average causal Effect）

此方法用于探究某干预方法（treatment A）对群体平均结果的影响。在统计学中，我们通常通过控制变量法来研究因素对群体的影响，即在保持其他因素不变的情况下，改变研究因素，然后观察群体的变化，以确定因素与群体之间的关系。然而，许多研究因素在现实中无法直接变化，例如探究某药物对某种疾病的效果。在这种情况下，我们无法同时观测病人群体在服药与不服药的状态。

平均因果效应通过建立因果模型（Causal Model）来研究。在这个模型中，我们可以设置干预方法（treatment A）的值分别为 1 和 0，即让目标变量（Y）在接受干预和不接受干预的情况下分别计算。通过这个操作，我们可以获得两个干预分布（Interventional Distribution）：Y|do(treatmentA:=1) 和 Y|do(treatmentA:=0)。计算这两个分布的均值，然后求差，即 E[Y|do(treatmentA:=1)]−E[Y|do(treatmentA:=0)]，我们将这个均值差称为平均因果效应（Average Treatment Effect 或 Average Causal Effect）。

2.6 干预分析（intervention analysis）

此方法与平均因果效应（Average Treatment Effect 或 Average Causal Effect）的计算过程基本一致，不同之处在于，我们在干预分析后直接给出两个干预分布（Interventional Distribution）作为结果呈现。

在干预分析（Intervention Analysis）中，我们有两种干预方法：原子干预（Atomic Intervention）和移位干预（Shift Intervention）。

原子干预（Atomic Intervention）：将干预因素设定为一个固定值，使群体中每个个体的干预因素都相同。这种方法可以消除干预因素对目标变量的影响。

移位干预（Shift Intervention）：将干预因素增加一个设定的值，使群体中每个个体的干预因素都增加相同的值。这种方法可以用来观察干预因素对目标变量的影响。

2.7 分布变化归因（attributing distributional change）

此方法旨在研究不同时间段内，两组数据是否发生了分布变化，并识别造成这种变化的因素。例如，某公司第二季度利润相比于第一季度下降了，我们需要找出导致利润下降的因素。

为解决这个问题，我们首先基于核的独立性测试（Kernel-based Conditional Independence Test）和因果图，检查两个时间段的数据是否分布一致。如果某些属性的分布一致，则新时间段的因果机制可以继承旧时间段的因果机制；如果某些属性的分布不一致，则需要重新判断并拟合新的因果机制。接下来，将两个数据集分别建立为可逆结构因果模型（Invertible Structural Causal Model），将这两个模型按时间命名为旧因果模型和新因果模型。根据这两个因果模型生成样本数据，并比较这两个样本数据的分布。随后通过 KL 散度（Kullback-Leibler Divergence）计算出两个分布的差异性，从而获得分布变化得分。最后，通过 Shapley 方法将分布变化得分分配给各个影响因素，从而得到所有因素的分布变化得分。

2.8 反事实分析（counterfactual analysis）

此方法用于研究在研究群体中某个个体的影响因素发生改变后，目标变量的变化。这在个体因果效应的估计、个性化决策、政策评估等领域有重要作用。例如，某个人的胆固醇含量很高（$Z = 10$），最近几个月他没有服用任何药物（$x = 0$），如果他每天服用 5 克药物（$x = 5$），他的胆固醇含量（$Z = ?$）会发生怎样的变化?

为了解决这个问题，我们首先基于研究群体建立可逆结构因果模型（Invertible Structural Causal Model）。然后，从研究群体中挑选出要研究的个体，对该个体中的某个或多个影响因素进行修改。接着，我们可以观察这些修改后该个体的其他因素及目标变量的变化。一、根据录入的数据以及链接关系，建立因果模型。二、根据因果模型以及链接关系，由之前的数据集，获得录入数据的噪声数据。该噪声数据每一条都对应录入数据中的每一条样本。三、导入改变后的数据样本，结合该数据样本的噪声数据以及因果模型，生成新的反事实数据。修改后的因素，也会导致后续它链接的因素数据发生变化。

2.9 异常归因（anomaly attribution）

此方法旨在探究哪些上游因素对目标节点异常评分的贡献度最大。例如，在研究某公司收入情况时，如果发现某条数据的收入异常高，我们需要确定导致这一异常的因素。

为解决这个问题，我们首先基于研究群体建立可逆结构因果模型（Invertible Structural Causal Model），并利用该模型确定目标变量的分布。然后，将观察到的异常个体与目标变量的分布结合，计算异常个体的异常得分（anomaly score）。最后，使用 Shapley 方法将异常得分分配给各个影响因素，从而得到各因素对异常个体的贡献得分。一、首先根据数据以及链接关系，生成因果模型。二、与内在因果分析相同，根据已有数据分布和因果模型，生成各节点的噪声。这里的噪声是指的值的误差以及波动，我们通过这些噪声来评估异常得分以及异常的成因。三、通过使用异常评分计算式：-log（P（g（X）>= g(x)））来获得总利润因素的异常得分，其中 g(x) 表示一个异常评分函数，他的作用是根据数据的特征（维度和分布形式）获得给定样本的异常程度，我们这里默认使用高斯混合模型（Gaussian Mixture Model, GMM）作为异常评分函数。其中 X 表示为目标因素，即总利润的正常数据分布。x 表示目标因素的异常数据。简单理解这个计算式，就是样本 x 越离群、越罕见，这个公式的值，即异常评分就越高。四、使用 shapley 值法，根据生成的异常分数，算出每个因素对总利润异常的贡献度。

2.10 因果发现（causal learn）

当用户不确定所研究项目的因果结构时，可以选用因果发现功能，让其根据数据，自动生成因果模型。因果发现功能不仅仅是发现变量之间的相关性，更重要的是确定哪些变量是原因，哪些变量是结果。因果发现功能主要分为以下方法：基于约束的因果发现方法、基于分数的因果发现方法、基于约束函数的因果发现方法、隐性因果表征学习、基于排列的因果发现方法。

2.10.1 基于约束的因果发现方法

基于约束的方法大多数是在经验联合分布上测试条件独立性，来构造一张反映这些条件独立性的图。通常会有多个满足一组给定的条件独立性的图，所以基于约束的方法通常输出一个表示某个 MEC（边缘计算）的图（例如，一个 PAG）。

2.10.1.1 Perform Peter-Clark（PC）

PC 算法是一种基于约束的因果发现方法，通过逐步执行条件独立性检验构建因果图骨架（删除无关边），随后利用碰撞结构（V-结构）和方向传播规则确定边方向，最终输出部分有向无环图（PDAG）。

该算法支持连续（如 Fisher's Z 检验）和离散数据（如卡方检验），默认假设变量服从多元高斯分布，但对非高斯数据可通过核方法扩展，其结果为马尔可夫等价类，适用于中等维度数据，但对高维场景计算复杂度较高，且依赖条件独立性检验的准确性。

计算步骤：

1. 从一个完全无向图开始，通过条件独立性（d-separation）测试逐步去除边来识别骨架。

2. 条件独立性测试用来评估 $X_i \perp\!\!\!\perp X_j|S$, 其中条件集合 S 可能为空。

3. 之后，如果骨架中给存在一条 $X_i - X_k - X_j$ 这样的路径满足 $X_i \perp\!\!\!\perp X_j$ 但是 $X_i \not\perp\!\!\!\perp X_j|X_k$, 那么我们可以知道这条路径事实上一定是一个 **v-结构（对撞结构）**，可以确定变得方向为 $X_i \to X_k \leftarrow X_j$。

4. 在确定了所有 **v-结构（对撞结构）**之后可以尝试给图中剩余的边确定方向，同时保持图的无环性，并尽量避免产生新的 **v-结构（对撞结构）**。

2.10.1.2 Fast Causal Inference (FCI)

FCI 算法是 PC 算法的改进，它通过验证初始骨架中邻接的两个顶点在给定 Possible-D-SEP 集下的条件独立信息来解决潜变量和选择变量带来的影响。首先，FCI 算法利用条件独立性测试来确定变量之间的初始关系，形成一个可能的因果结构集合。然后，它通过分析这些关系的特点，进一步筛选和细化因果结构，识别出那些在不同条件下保持稳定的因果关系。

计算步骤：

1. 构造一个非定向边构成的完全图，对于顶点 X_i 和 X_j，在给定 X_i 的邻接子集 M 的条件下，对俩个顶点做条件独立检验，如果条件独立关系成立，则移除 X_i 和 X_j 之间的边，并且把相应的邻接子集 M 保存到分离集 $sepset(X_i, X_j)$ 中。第一步得出的骨架式最终骨架的超集

2. 对于开放三元组 $< X_i, X_j, X_k >$,若 X_j 不在 $sepset(X_i, X_k)$ 和 $sepset(X_k, X_i)$ 中，则把此三元组定向成 **v–结构（对撞结构）**

3. 对于 X 中所有顶点 X_i,找出它的 $pds(X_i, \cdot)(Possible-D-sep)$。然后对于所有 X_i 的邻接顶点 X_j, 检验 $X \perp X_j|(Y \cup S)$ 是否存在, $Y \subseteq pds(X_i, \cdot)\ X_i, X_j$。若存在则移除两顶点之间的边,并把相应的邻接子集 M 保存到分离集 $sepset(X_i, X_j)$

和 $sepset(X_j, X_i)$ 中。这样就破坏了图的骨架，需要重新定向 v–结构。所以最后把所有的边重新改为非定向边。

4. 重复第二步，重新定向 v–结构。

2.10.2 基于分数的因果发现方法 (Score-based causal discovery methods)

基于分数的因果发现方法将因果发现问题转化为一个优化问题，假设存在一个潜在的因果结构空间，其中每个因果结构都对应着一个分数，该分数用于衡量该结构与观测数据的拟合程度或对数据生成过程的解释能力。通过搜索这个结构空间，找到使评分函数值最优的因果结构，作为对真实因果关系的估计。

2.10.2.1 GES with the BIC score or generalized score

GES 是一种启发式搜索算法，旨在从一组可能的结构中找到最优的因果结构。它从一个初始结构开始，通过迭代地进行局部更改（例如添加或删除边）来提高结构的分数。在此情境下，贝叶斯信息准则（BIC）分数被广泛使用。该分数在因果模型对数据的拟合程度与模型复杂度的惩罚之间进行平衡，有助于防止过拟合.

计算步骤：

1. 前向搜索（Forward Search）：初始化一个空图（无边的图），在每一步，尝试添加一条边（$X \to Y$ $X \leftarrow Y$），并计算评分函数的变化。随后选择使评分函数最大化的操作（即添加使评分增加最多的边）。重复上述步骤，直到无法通过添加边进一步提高评分。

2. 后向搜索（Backward Search）：从前向搜索得到的图开始，在每一步，尝试删除一条边（$X \to Y$ $X \leftarrow Y$），并计算评分函数的变化。选择使评分函数最大化的操作（即删除使评分增加最多的边）。重复上述步骤，直到无法通过删除边进一步提高评分。

常用的评分函数：

BIC（Bayesian Information Criterion）：

$BIC = \log P(D \mid G) - \frac{k}{2} \log n$

其中，D 是数据，G 是因果图，k 是模型参数的数量，n 是样本量。

BIC 倾向于选择较简单的模型（惩罚复杂度）。

BDeu（Bayesian Dirichlet Equivalent Uniform）：

基于贝叶斯方法的评分函数，适用于离散数据. 对于一个贝叶斯网络结构 G 和观测数据 D，BDeu 评分的公式为：

$$\mathrm{BDeu}(G|D) = \prod_{i=1}^{n} \prod_{j=1}^{q_i} \frac{\Gamma(\alpha_{ij})}{\Gamma(\alpha_{ij} + N_{ij})} \prod_{k=1}^{r_i} \frac{\Gamma(\alpha_{ijk} + N_{ijk})}{\Gamma(\alpha_{ijk})}$$

n: 表示贝叶斯网络中的变量的数量
q_i: 变量 X_i 的父节点所有可能取值组合的数量。比如，若变量 X_i 有两个父节点，每个父节点有 2 种取值，那么 q_i=2×2=4。
r_i: 变量 X_i 可能取值的数量。
α_{ij}: 是 Dirichlet 先验分布的超参数，在 BDeu 中，它的计算方式为 $\alpha_{ij} = \frac{\alpha}{q_i}$ ，其中 α 是等价样本量（equivalent sample size）。
α_{ijk}: 同样是 Dirichlet 先验分布的超参数，在 BDeu 里，$\alpha_{ijk} = \frac{\alpha}{q_i r_i}$
N_{ij}: 在数据 D 中，变量 X_i 的父节点取值为第 j 种组合的样本数量。
N_{ijk}: 在数据 D 中，变量 X_i 的父节点取值为第 j 种组合, 且 X_i 取值为第 k 种状态的样本数量。
$\Gamma(\cdot)$: 是伽马函数,对于正整数 n,$\Gamma(n) = (n-1)!$; 对于非整数 x,$\Gamma(x) = \int_0^\infty t^{x-1}e^{-t}dt$.
该公式是对网络中每个变量及其父节点的不同取值组合的概率进行连乘。每个变量的得分是由其在不同父节点取值组合下的概率贡献累积得到的。等价样本量 α 起到平衡先验信息和观测数据的作用。当 α 较大时，先验信息对评分的影响更大；当 α 较小时，观测数据的影响更显著。

2.10.2.2 Exact Search：

使用动态规划（DP）或 A* 搜索寻找最优图

它通过系统地遍历所有可能的因果结构空间，对每个结构依据特定的评分函数或准则进行评估，如贝叶斯评分、似然评分等，然后从中挑选出得分最优的因果结构。这种方法虽然能保证找到全局最优解，但由于因果结构空间随着变量数量的增加呈指数级增长，计算复杂度极高，在实际应用中，对于大规模变量系统往往难以实施，通常适用于变量数量较少、对因果结构准确性要求极高的场景。

它采用了两种搜索策略：动态规划（Dynamic Programming，DP）和 A* 搜索。这些方法通过系统地探索所有可能的图结构，寻找全局最优解。

精确搜索的工作原理：

1. 动态规划（DP）：DP 方法通过分解问题，将复杂的图结构学习任务转化为多个子问题的求解，从而高效地找到最优解。
2. A* 搜索：A* 搜索是一种启发式搜索算法，它在搜索过程中利用启发式函数评估每个节点的潜在最优性，从而优先探索最有可能导致最优解的路径。

3 案例

3.1 案例背景

本教程中使用的示例数据集是超市运营数据集。本数据集一共有 2000 条数据，10 列变量。

表 1: 超市运营数据集

列名	描述	数据类型
MaterialCosts	材料费用	float
RunningCosts	运营费用	float
Advertising	广告费用	float
ProductPrice	产品售价	float
Weather	天气	float
Footfall	客流量	float
Expenses	成本	float
Sales	销售额	float
Profit	利润	float
Index	序号	float

在本教程中，假设您是一名超市老板。您最近在查看财务报表时，发现利润同比下降了，您希望使用因果的方法找到影响利润下降的主要因素，并且希望找到一个恰当的改进方法。

3.2 创建项目

进入主页后，点击开始使用或者登录进行登录。默认用户名与密码为 admin。

图 65: 开始使用

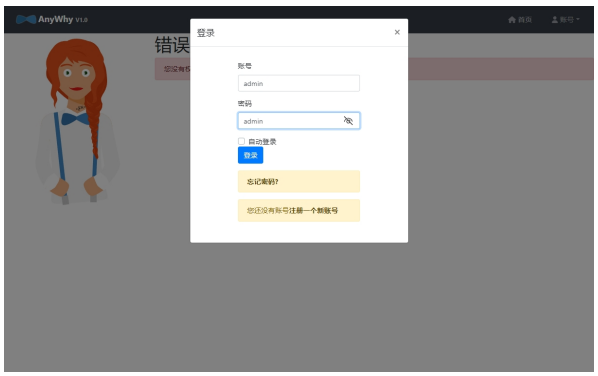


图 66: 登录界面

进入项目管理页面后，点击“创建新 Project”创建新项目。

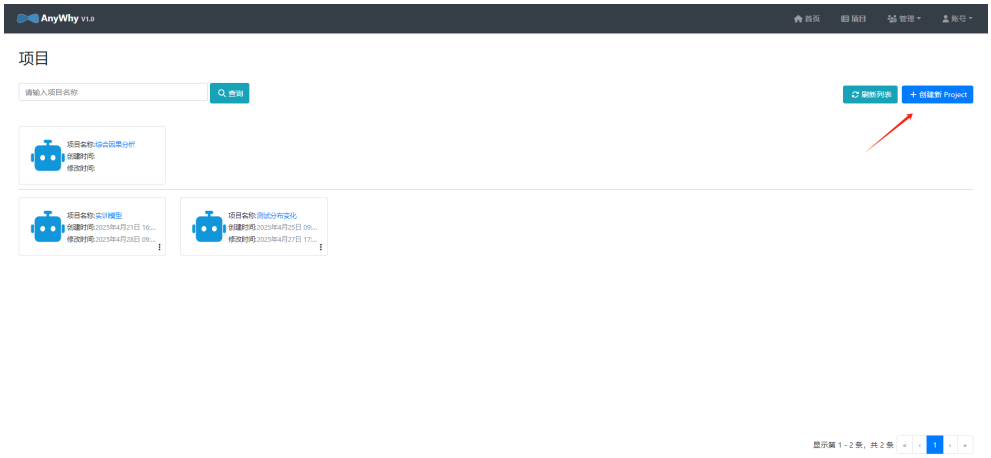


图 67: 创建新项目

填入项目名称后，进入项目编辑页面，项目创建完成。

图 68: 创建项目完成

3.3 导入数据、模型建模

3.3.1 手动建立模型

点击“导入数据”按钮，选择 excel 或 csv 文件进行数据导入。

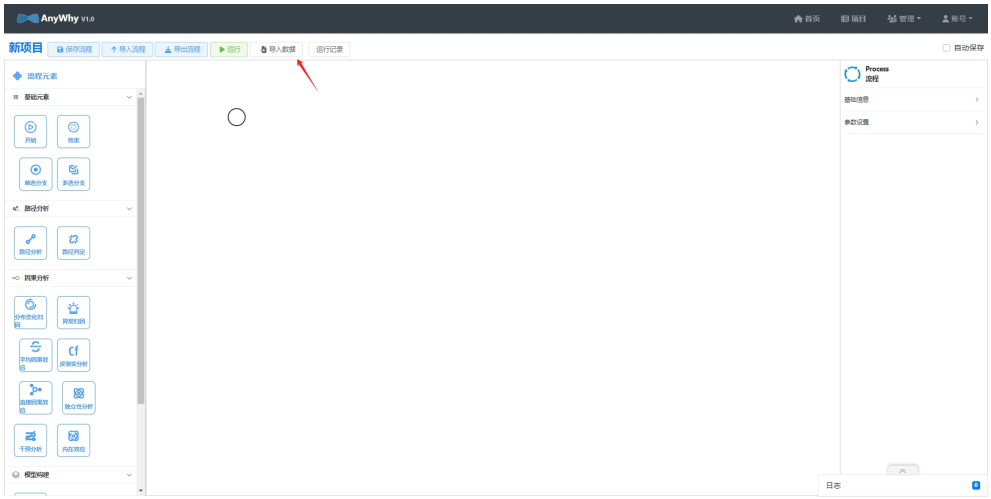
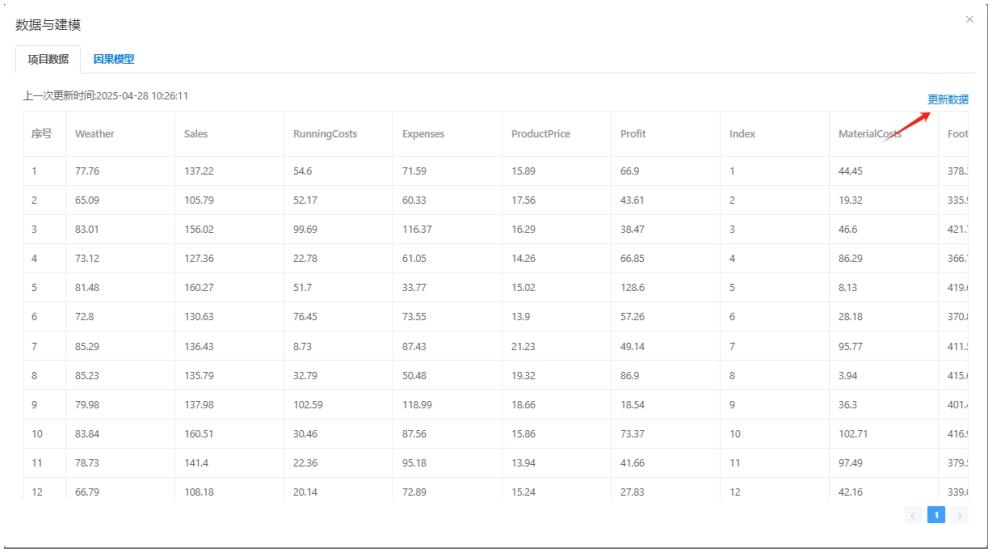


图 69: 导入数据

导入完成后，“导入数据”按钮变为“查看数据”按钮，点击可以查看已导入的数据，如果需要更改数据，点击查看数据界面下的“更新数据”按钮



序号	Weather	Sales	RunningCosts	Expenses	ProductPrice	Profit	Index	MaterialCosts	Foot
1	77.76	137.22	54.6	71.59	15.89	66.9	1	44.45	378.
2	65.09	105.79	52.17	60.33	17.56	43.61	2	19.32	335.
3	83.01	156.02	99.69	116.37	16.29	38.47	3	46.6	421.
4	73.12	127.36	22.78	61.05	14.26	66.85	4	86.29	366.
5	81.48	160.27	51.7	33.77	15.02	128.6	5	8.13	419.
6	72.8	130.63	76.45	73.55	13.9	57.26	6	28.18	370.
7	85.29	136.43	8.73	87.43	21.23	49.14	7	95.77	411.
8	85.23	135.79	32.79	50.48	19.32	86.9	8	3.94	415.
9	79.98	137.98	102.59	118.99	18.66	18.54	9	36.3	401.
10	83.84	160.51	30.46	87.56	15.86	73.37	10	102.71	416.
11	78.73	141.4	22.36	95.18	13.94	41.66	11	97.49	379.
12	66.79	108.18	20.14	72.89	15.24	27.83	12	42.16	339.

图 70: 更新数据

点击因果模型选项

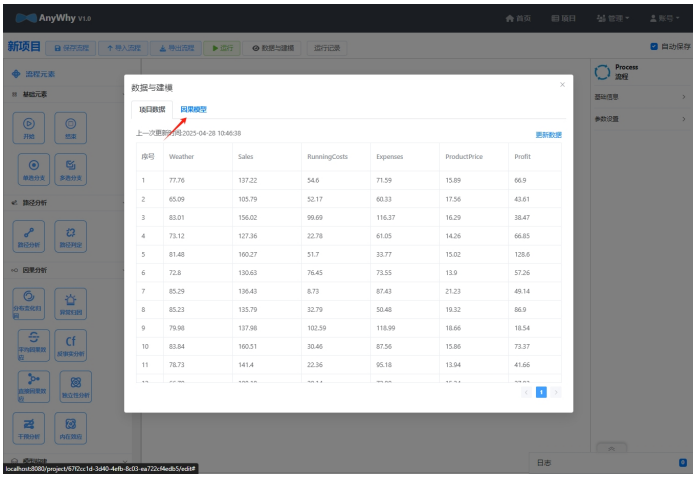


图 71: 点击因果模型选项

这里可以创建因果模型

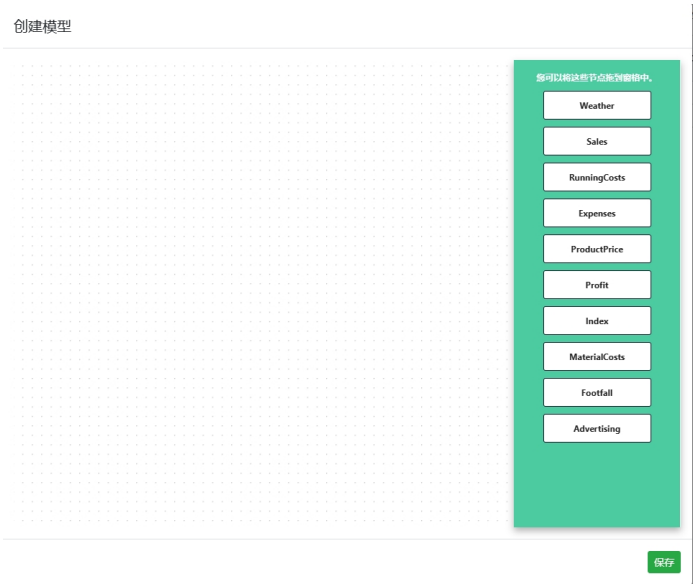


图 72: 创建模型界面

拖拽右侧节点到画布上，使用鼠标可以将节点连接。使用**退格键**可以将节点或者连接线删除。

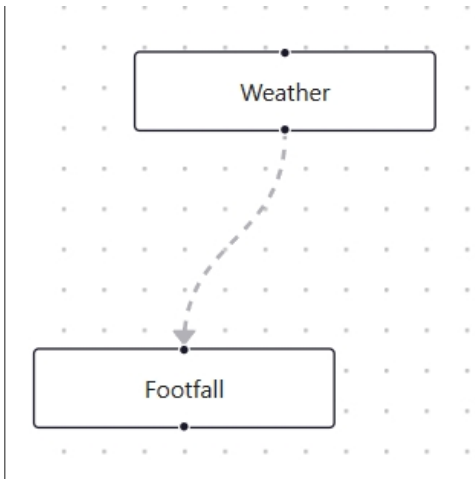


图 73: 节点连接

借此，我们可以构建因果模型。

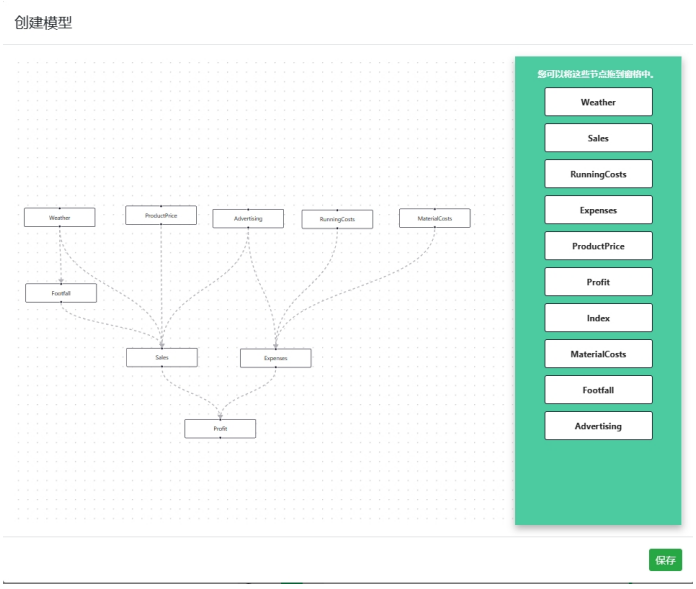


图 74: 因果模型建立完成

随后，我们需要点击非根节点，即 Profit，Sales，Expenses，Footfall，为其设定连接机制，这里我们都设定为 linear。**注意，为根节点设定不会影响最终计算结果**

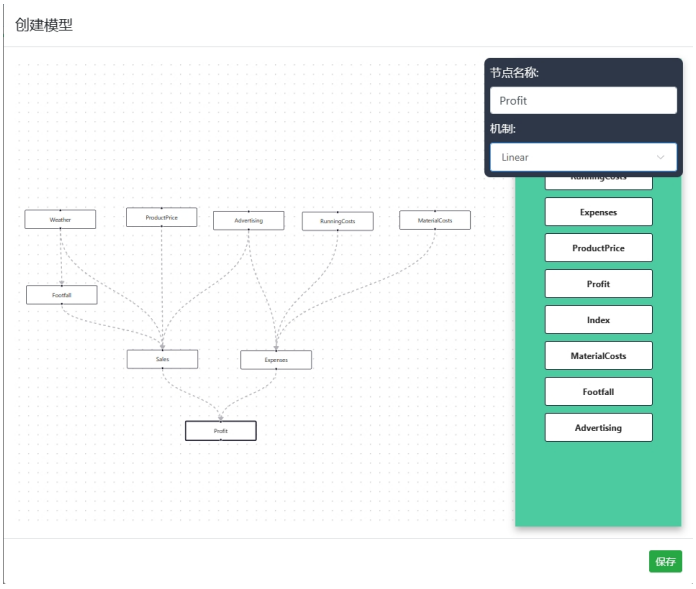


图 75: 设定连接机制

当然，不知道连接机制的时候，我们可以设置自动设置因果机制，让其自动匹配合适的连接机制。

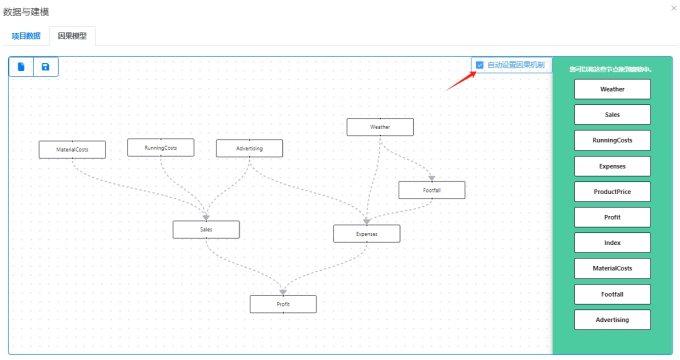


图 76: 自动设置因果机制

右上角可以保存因果模型。

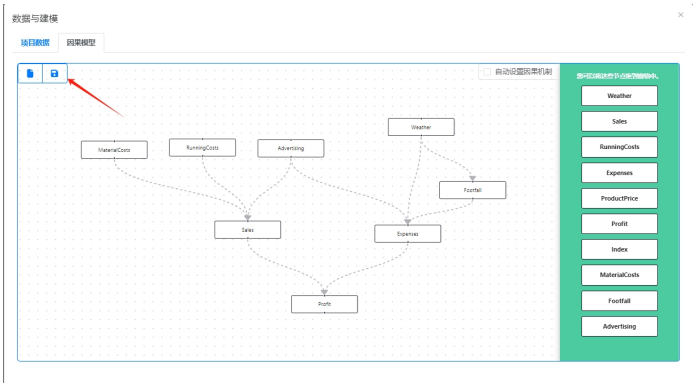


图 77: 保存因果模型

建立完成后点击保存保存模型。
或者我们也可以使用果模型建模节进行手动因果建模。
从节点处拖入“因果模型构建”节点

图 78: 拖入模型构建节点

点击“流程开始”节点，选择箭头，与“因果模型构建”节点相连

图 79: 选择箭头

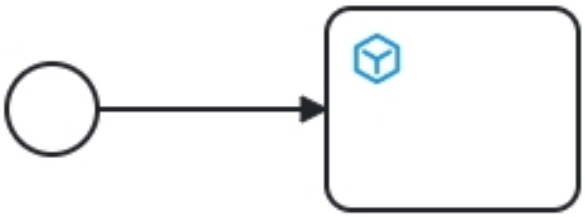


图 80: 与模型构建相连

点击“运行”按钮，将界面由编辑状态转换为运行状态，随后点击位于“流程开始”节点上的“运行”图标

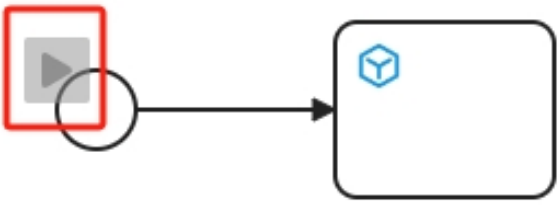


图 81: 运行图标

运行到“因果模型构建”节点时，弹出“创建模型”界面，录入数据的表头将会出现在界面右侧。

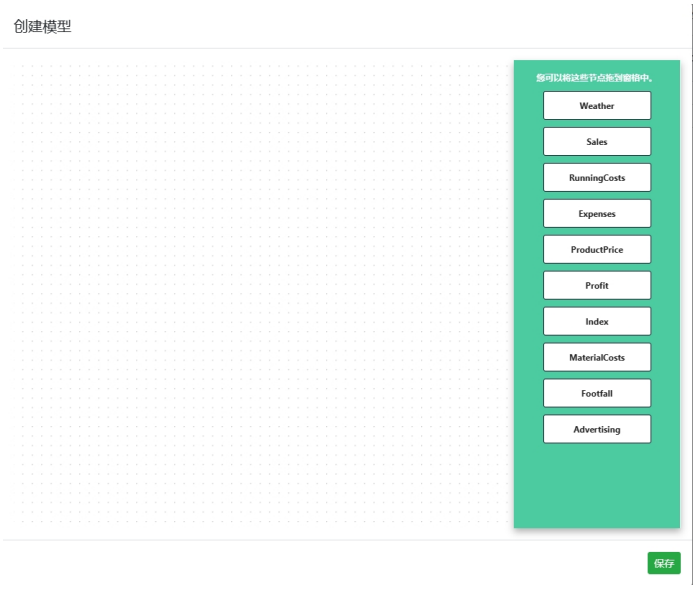


图 82: 创建模型界面

在这里的因果构建和上述相同，但是无法自动拟合连接机制。点击保存即可使流程继续。

3.3.2 使用因果发现生成因果模型

点击因果模型左上角“生成因果模型按钮”

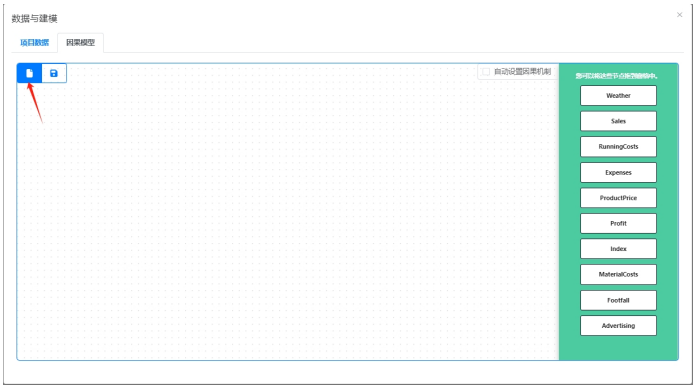


图 83: 生成因果模型按钮

因果发现参数选择，选择“基于分数”下的“ges 算法“，点击生成。

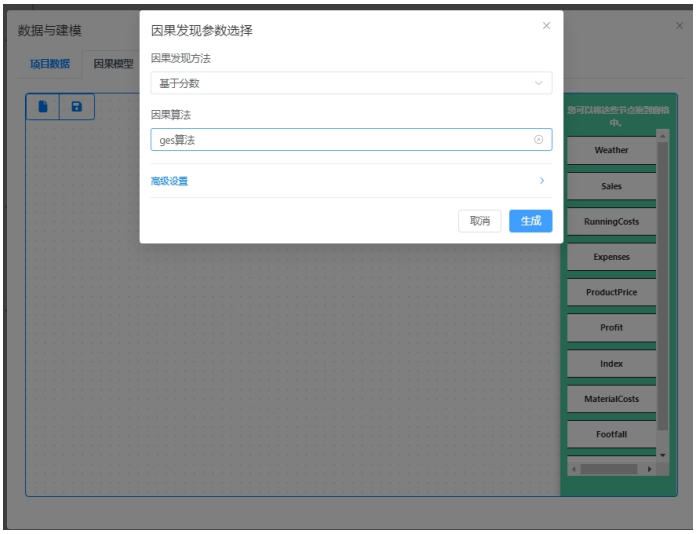


图 84: 因果发现参数选择

获得自动生成的因果模型

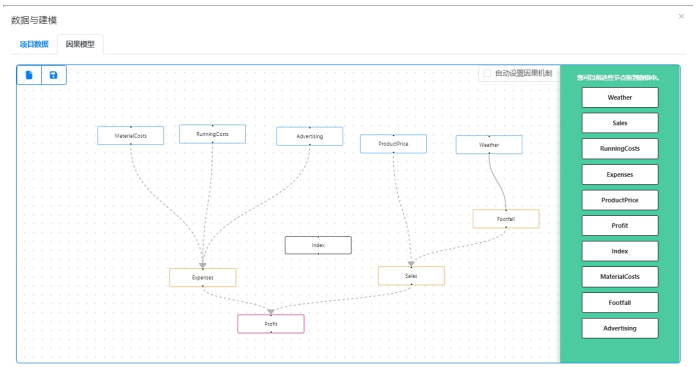


图 85: 自动生成的因果模型

将 Index 去掉，将无箭头线删除，换为箭头线；将缺失的线添加上。并且将连接机制设置好。最后点击保存。

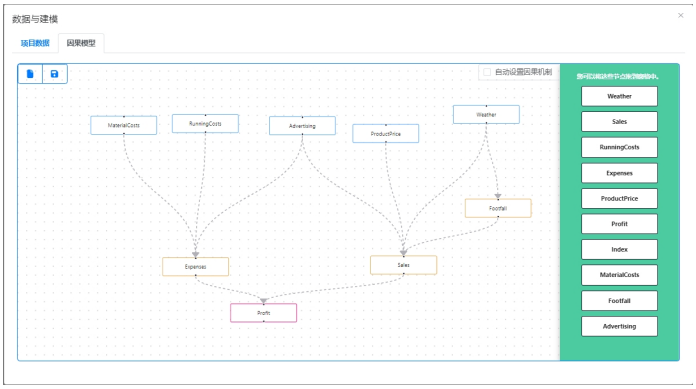


图 86: 自动生成的因果模型调整

在下面我们使用的因果模型节点建模进行操作。如果使用”导入数据“中的”因果模型“进行的建模，我们无需拖动因果建模节点到操作区，直接进行分析即可。

3.4 独立性分析

将判断节点和独立性分析节点拉到操作区，因果模型建模节点和判断节点相连，判断节点和独立性分析节点相连。

图 87: 连接判断节点和独立性分析节点

点击判断节点和独立性分析的箭头，在右边栏“名称”处填入“独立性分析”，箭头上方会出现“独立性分析”字样。这样，在运行时，经理判断节点时，选项框中将出现“独立性分析”选项。

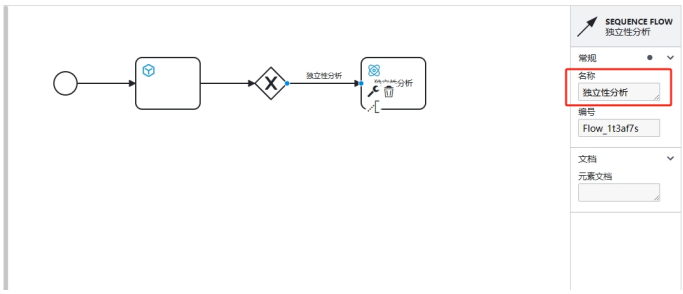


图 88: 箭头名称

点击“运行”按钮，点击位于“流程开始”节点上的“运行”图标，保存因果模型后进行计算，计算完成后，“独立性分析”节点上方出现 Finished

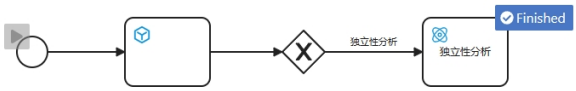


图 89: 独立性分析运算完成

点击右上角日志即可查看结果

查看结果发现广告费用与销售额相互独立，两个之间不应该有关系。

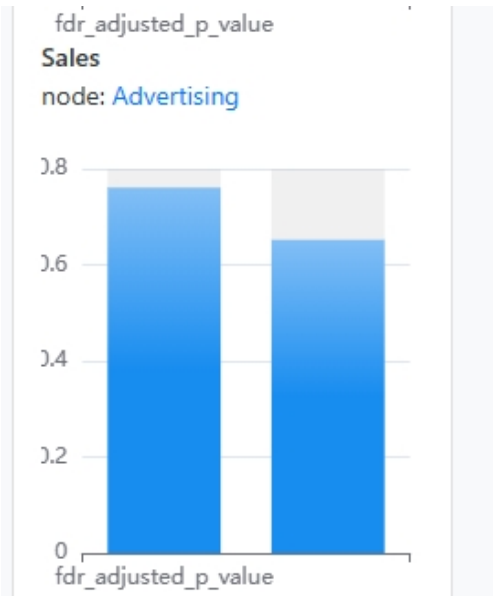


图 90: sales 与 advertising 独立性分析结果

修改因果图，将 sales 和 advertising 之间连线删除。

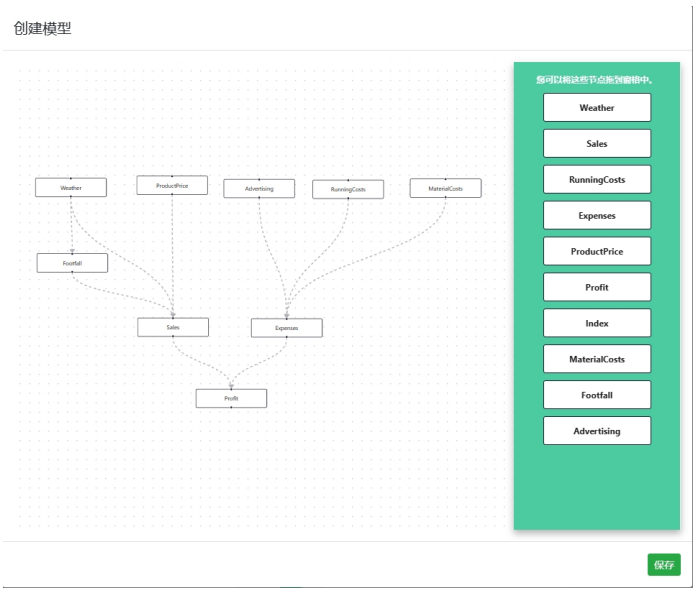


图 91: 修改后因果模型

3.5 路径分析

将“路径分析”节点和“路径是否关闭”节点加入流程。

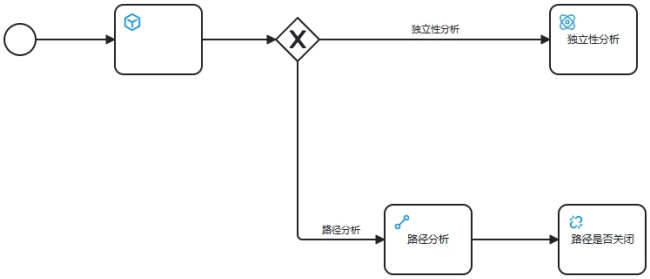


图 92: 添加路径分析

添加完成后，判断选择框会出现“路径分析”选项。

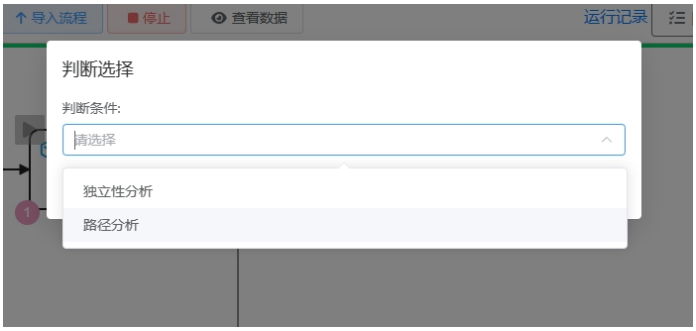


图 93: 判断选择

运行流程，处理变量填入 weather，目标变量填入 profit

图 94: 路径分析参数录入

在路径是否关闭处填入 footfall

图 95: 填入 footfall

查看在运行日志中的结果，发现 weather 到 profit 有两条路径，点击路径结果将在图中高亮显示。选择 footfall 没有阻断任何后门路径。

图 96: 路径分析结果

图 97: 路径是否关闭结果

3.6 直接因果效应

将“直接因果效应”节点加入流程。

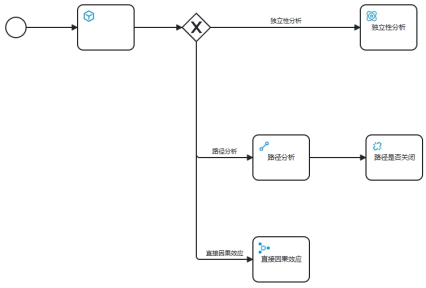


图 98: 加入直接因果效应节点

运行流程，这里我们不选择使用计算置信区间以及高级参数。

图 99: 直接运行直接因果效应

运行完成后，查看结果，发现成本影响总利润程度要大于销售额对总利润的影响。

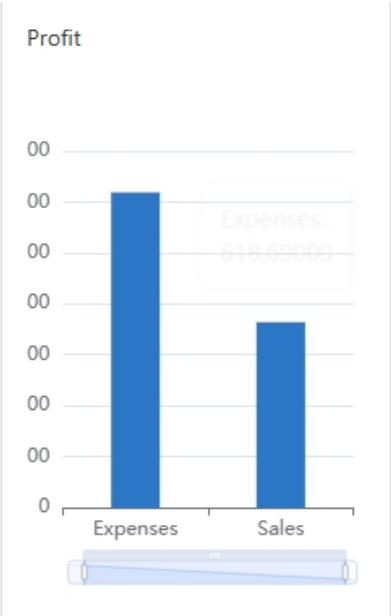


图 100: 直接因果效应结果-1

影响总成本的三个因素程度基本一样。

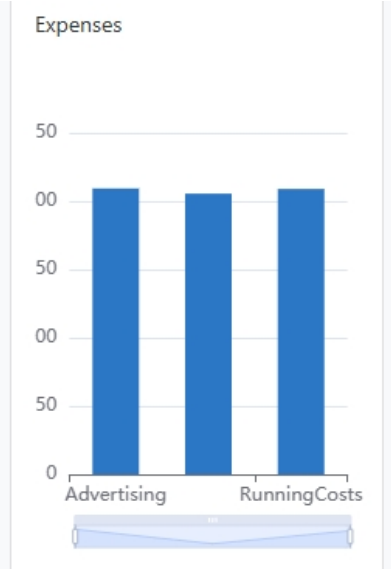


图 101: 直接因果效应结果-2

影响销售额的三个因素中，客流量影响最大。

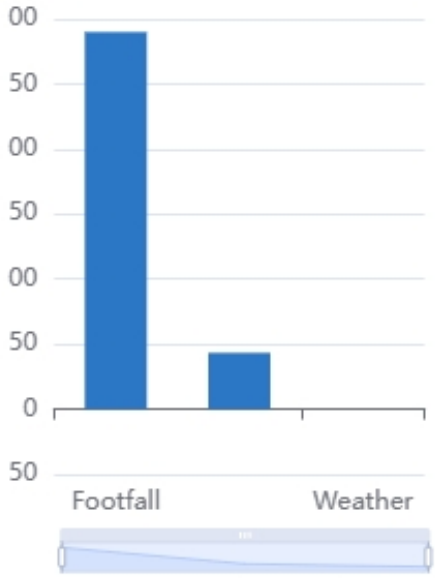


图 102: 直接因果效应结果-3

3.7 内在因果效应

将“内在因果效应”节点加入流程。

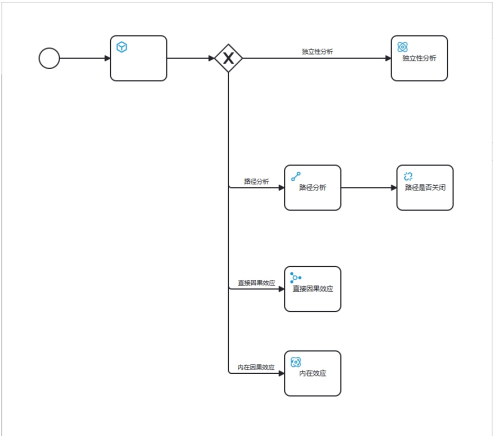


图 103: 加入内在因果效应节点

运行流程，这里我们不选择使用计算置信区间以及高级参数。

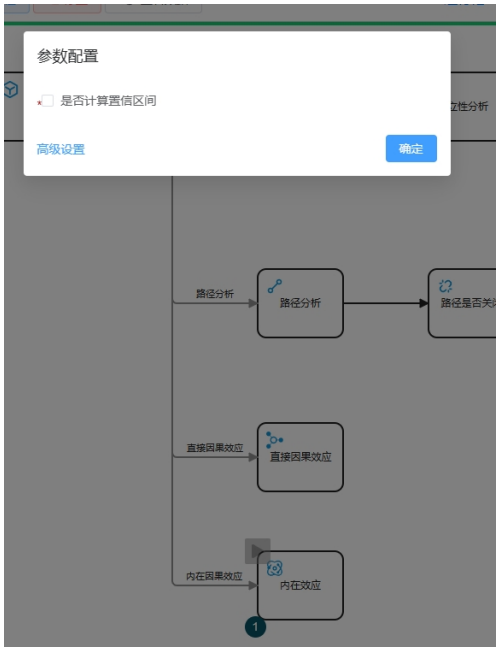


图 104: 直接运行内在因果效应

运行完成后，查看结果，发现影响总利润最大的因素是天气；除了天气外，材料费、运营费和广告费影响利润的程度相当。

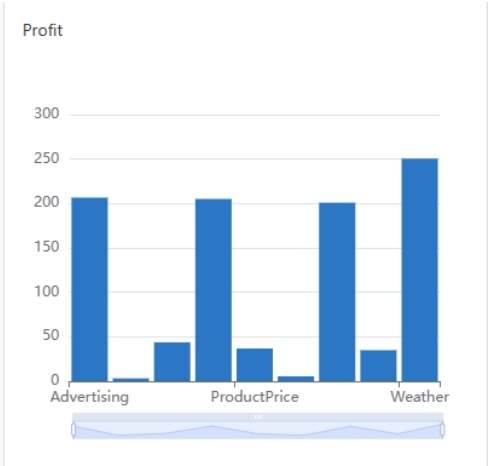


图 105: 内在因果效应结果

3.8 平均因果效应

使用平均因果效应，我们可以大致查看改变一个因素后，目标变量平均变化多少。将平均因果效应加入流程。

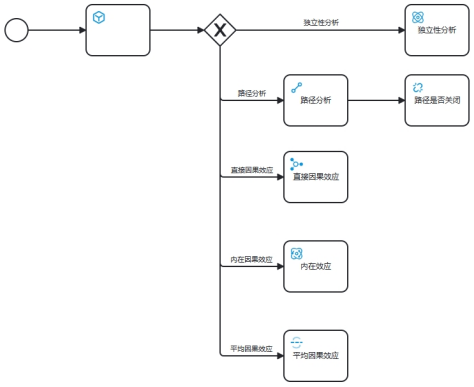


图 106: 平均因果效应加入流程

运行流程，这里我们可以选择改变运营成本。观察当运营成本平均 70 变化到平均 100 时，利润变化情况。

图 107: 平均因果效应填入参数

结果我们可以看到，利润平均变化为-15.21312

图 108: 平均因果效应结果

3.9 干预分析

将干预分析节点加入流程。

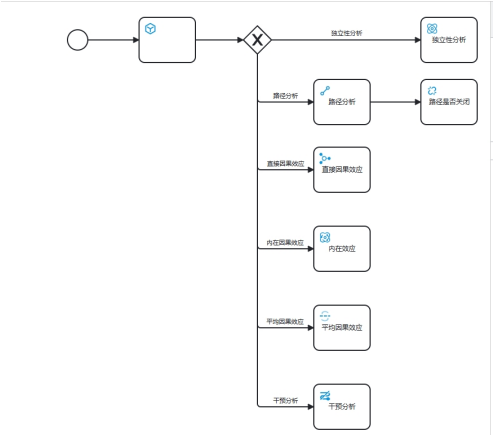


图 109: 干预分析加入流程

运行流程，这里我们可以选择干预运营成本。使用移位干预，观察当运营成本平均增加 30 时，利润数量分布变化情况。

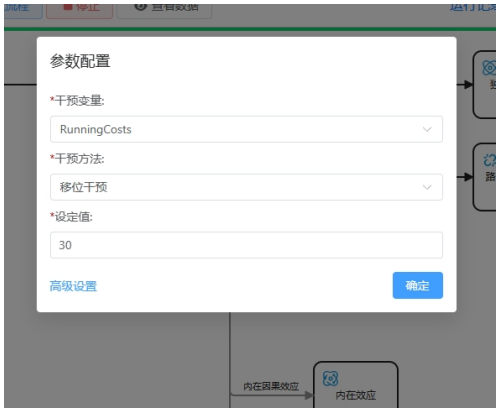


图 110: 配置干预分析

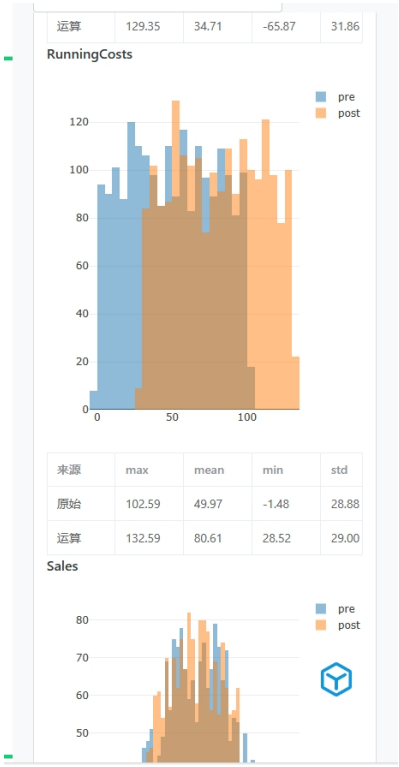


图 111: 干预分析结果

3.10 反事实分析

将反事实分析节点加入流程

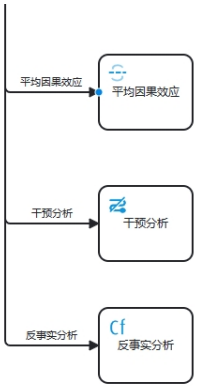


图 112: 反事实分析加入流程

点击运行，我们将第三条的广告费设定为 0，查看不投入广告费的情况下，利润将会增加还是减少。

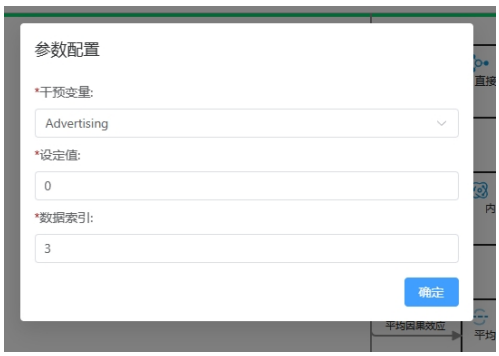


图 113: 反事实分析填入参数

查看结果，发现利润从之前的 66.85 变化到了 73.67.

图 114: 反事实分析结果

3.11 异常归因

将异常归因节点以及反事实节点加入流程

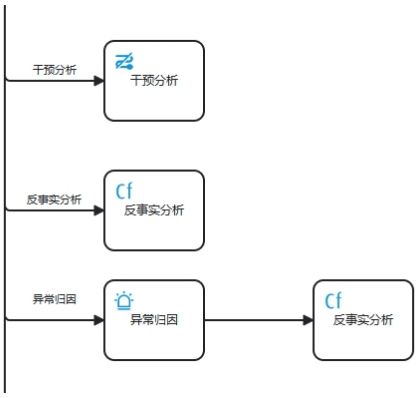


图 115: 异常归因以及反事实节点加入流程

点击运行，我们选择 ECOD 的异常检测方法来检测数据中 profit 的异常值

图 116: 异常归因填入参数

选择关注的异常数据索引

图 117: 异常数据选择

选择异常数据索引之后，在随后的反事实分析节点上，可以观察到异常数据索引的结果，我们可以参考结果填写反事实参数。

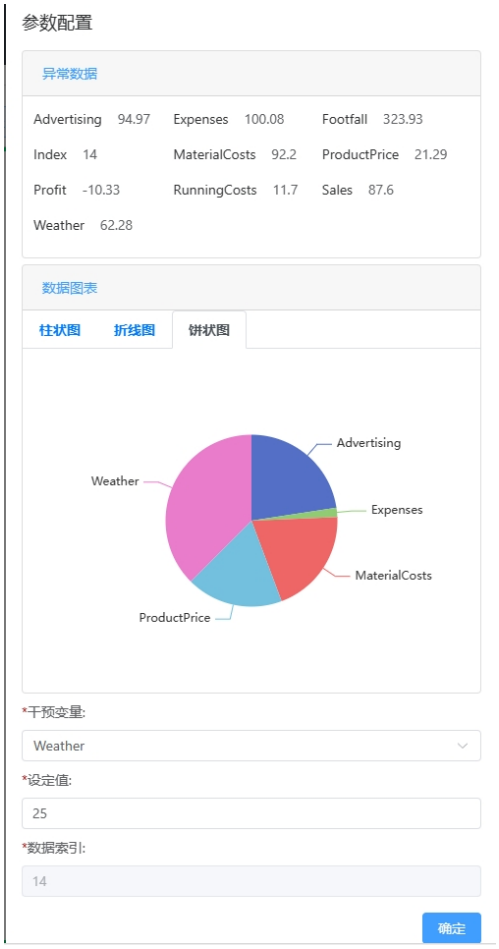


图 118: 异常归因后反事实分析参数填入

在运行日志中，我们可以看到反事实分析的结果，当 weather 从 62.25 变为 25 时，profit 从-10.33 变为了-98.95。亏损更多了。

图 119: 异常归因后反事实分析结果

3.12 分布变化归因

将分布变化归因加入流程

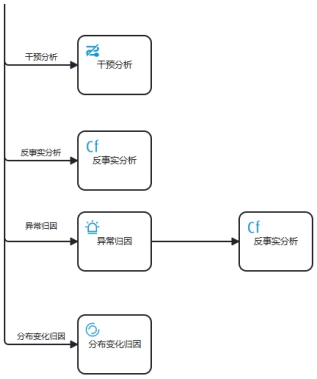


图 120: 分布变化归因加入流程

点击运行，我们将数据以索引为 1000 划分为两部分，第一部分为 1-1000，第二部分为 1001-2000。使用变化归因能探究这两个部分影响利润的变化原因。

图 121: 分布变化归因填写参数

查看结果，发现即使是影响利润最大的因素运营费用也只有 0.00168，这表示两个分布基本没有变化。

图 122: 分布变化归因结果