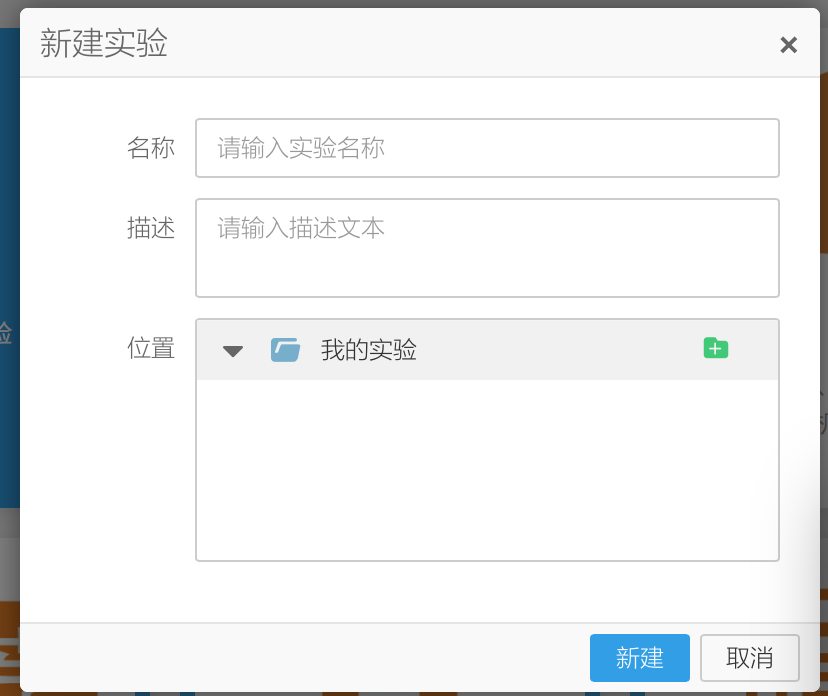
第一步：新建实验/已有实验的展示/模板实验

p1新建实验

1.名称

2.描述



p2展示已有实验

1.样式需要设计

2.包含元素：名称、描述、创建时间、数据源、选择的特征（可选）、使用的模型（可选）、模型的表现（可选）、状态

需要额外考虑的问题：同一个实验中更换了预处理方式、特征、模型等参数后，每次不同的模型表现，如何记录？是否作为同一次实验？

p2模板实验

1.样式类似已有实验

2.包含元素：除了上面的之外，可能需要在描述中添加较为详细的说，让其他人知道这个模板的作用

第二步：选择数据源

前端的使用上

p1.该类型节点没用输入桩，只有输出桩

p2.需要有字段表，可能的话需要有若干个真实数据可以看

p2.对于姓名、订单号、身份证、手机号、联系人等敏感信息，需要隐藏

p3.未来是否需要支持用户自己导入数据。因为可能存在进件、贷后变现数据以外的数据，而我们没有。尤其是具体到每个不同部门。

后台数据的整合

p1.示例用的江湖救急+学历/证书数据

p2.kg+三方+规则引擎+已知特征要计算好。需要定期把同一业务的数据整合好

p2.对于三方数据和规则引擎里面的数据，是拆分到每个字段去存，还是存一个json，然后每次去解析他们？需要权衡。

第三步：数据处理

p1.归一化

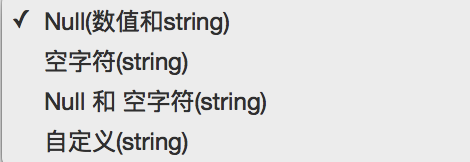
选择字段、保留原始列（可选）

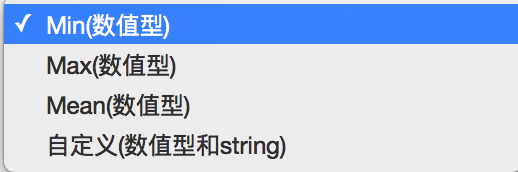
p1.标准化

选择字段、保留原始列（可选）

p1.缺失值填充

填充的字段、原值（Null、空字符、Null和空字符、自定义）、替换为





p1.类型转换

转换为double类型的列+转换为double异常时默认填充值；

转换为int类型的列+转换为int异常时默认填充值

转换为string类型的列+转换为string异常时默认填充值

保留原始列（可选）

描述：将任意类型特征转成string|double|int特征，并支持转换异常时的缺失值填充

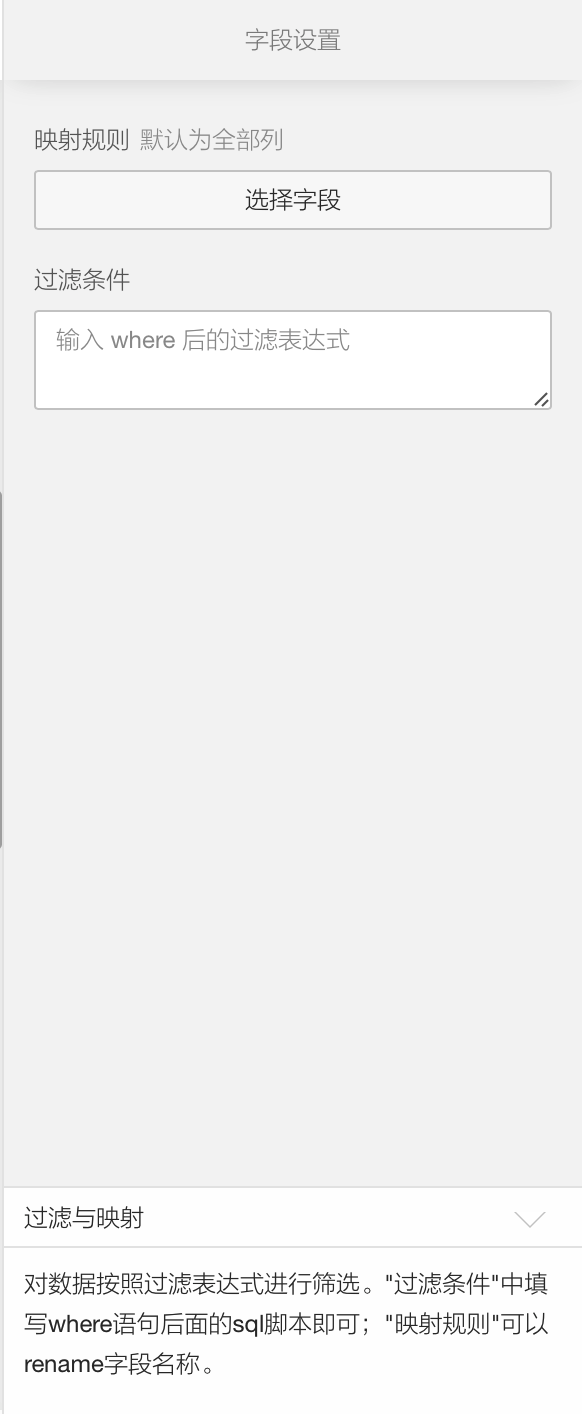


p2.过滤与映射

映射规则：选择字段

过滤条件：输入where后的过滤表达式

描述：对数据按照过滤表达式进行筛选。“过滤条件”中填写where语句后面的sql脚本即可；“映射规则”可以rename字段名称。



p2.拆分

切分比例（0，1）；

描述：按照比例拆分样本集，可以用于测试、验证

p2.随机采样

采样个数/采样比例（0，1） 二选一

放回采样（可选框）

随机数种子 正整数（默认为空）

描述：以随机方式生成采样数据，每次采样时各自独立的。

p2.分层采样

分组列

采样个数/采样比例（0，1） 二选一

随机种子值（可选）

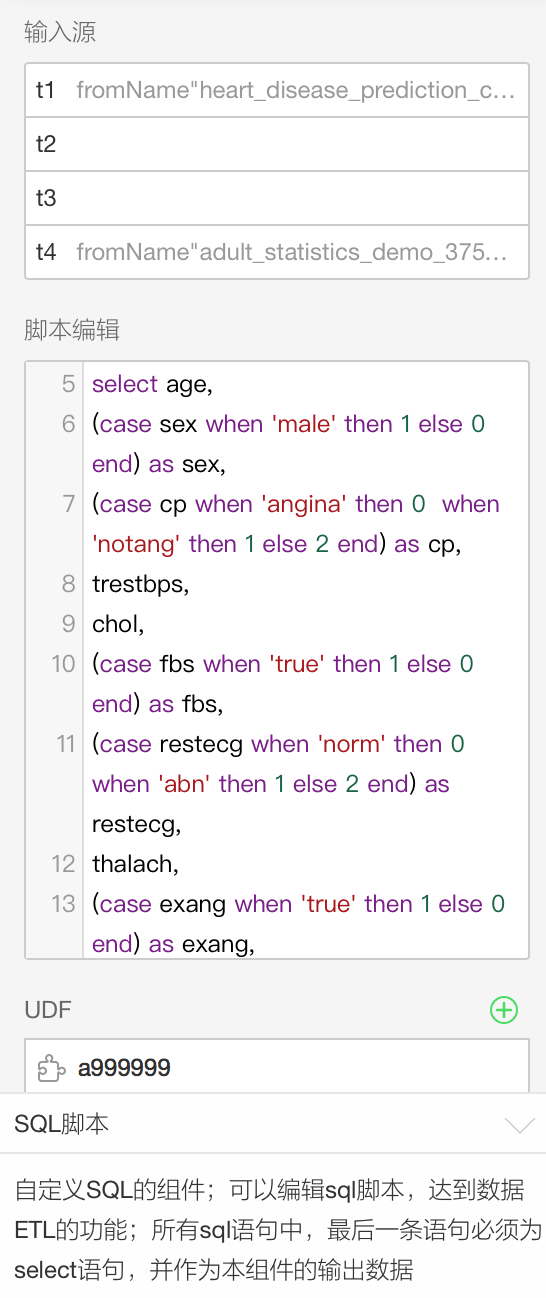
描述：根据指定的分组字段分层采样样本



p2.自定义SQL脚本

脚本编辑

UDF？创建自定义函数？意思是导入本地的代码？



p2.数据合并

即数据表的加宽，为了：

1.当运算出了新的特征/字段时，可以与原有的数据一起进行后面的计算。

2.配合用户导入自己的数据使用。

这样才能让复杂的运算处理得到的结果字段有用。

p3.支持运算表达式

可以更方便进行数据处理。尤其对于统计分析的用户

p3.支持多种语言的代码编写：python、java、sass、r等

第四步：特征选择

p1.左边栏的特征列表

p2.建立特征库。包含前端页面和后台的内容。需要的属性为：

1.业务属性：类型（int string、中英描述、取值范围、覆盖率、开发时间、开发人员、是否强特征、应用业务线、是否上线 、数据来源

2.分析属性：auc、ks、iv、psi（需要划分时间段）、覆盖率、特征之间相关性、比例、正负样本、4张图（单调性）

3.包含功能：

a.搜索、筛选（多维度，筛选覆盖率、表现值等）

b.个人库（收藏）+公共库

c.命名空间（层级关系、分组、标签）

d.使用说明

e.版本

f.权限

p2.添加新特征——新特征是用已有的字段进行运算的结果。

1.在特征库那边添加新特征。支持多语言。这应该算是公共特征。可以修改。

2.自定义的SQL脚本，可以定义为私人的特征。

第五步：模型选择

machine learning模型

p1.logistic regression需要参数：

1.训练特征列（需为double/int类型字段）

2.目标列

3.正类值

4.正则项（可选，None，L1，L2）

5.最大迭代次数

6.正则系数（可选）

7.最小收敛误差

p1.random forest需要参数

1.选择特征列（可选，默认为除label列、权重列外的所有列）

2.排除列（可选，但不可与特征列共存）

3.强制转换列（可选，bigint解析成categorical）

4.选择权重列

5.选择标签列



统计相关模型

p1.直方图/折线图：

1.输入时选择字段

2.字段

3.搜索框

4.步长

5.图

6.图下方的拖拽

7.与折线图切换

8.导出图片

p2.全表统计

1.输入列（默认全部列）

2.结果展示每一列的常用统计属性值（datatype，totalcount，count，missingcount，nancount，positiveinfinitycount，negativeinfinitycount，min，max，mean，variance，standarddeviation，standarderror，skewness，kurtosis，moment2，moment3，moment4，centralmoment2，centralmoment3，centralmoment4，sum，sum2，sum3，sum4）