# RiskModeler 使用文档

目录

[RiskModeler 使用文档 1](#_Toc46432152)

[1.概要 1](#_Toc46432153)

[2.开始项目 2](#_Toc46432154)

[3.导入数据集 2](#_Toc46432155)

[4.数据集处理 3](#_Toc46432156)

[5.交互式分组 3](#_Toc46432157)

[6. 模型评分 4](#_Toc46432158)

[7. 数据集打分 5](#_Toc46432159)

## 1.概要

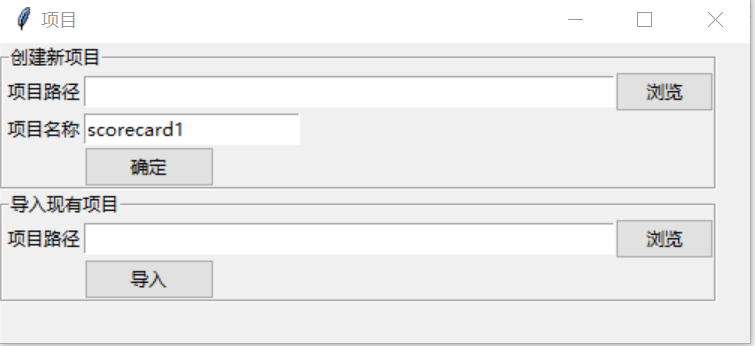
RiskModeler（RM）是一个风险评分卡可视化开发工具，主要参考SAS EM的模式实现评分卡开发流程的可视化。RM包括下面几个模块

1. 导入数据集（DATA）
2. 数据集处理
   1. 数据抽样(SAMPLE)
   2. 数据分区(SPLIT)
3. 交互式分组(IGN)
4. 评分卡
   1. 训练模型（SCR）
   2. 数据集打分（Scoring）
5. 导入模块

四大模块完成了数据集的导入，训练集验证集的分区，抽样，变量分箱，模型训练和数据集打分的功能。

（备注：用于GUI部分引用了Tkinter,因此在程序进行大量计算或者数据IO操作的时候，页面会显示无响应，请耐心等待。不要操作页面免得程序崩溃）

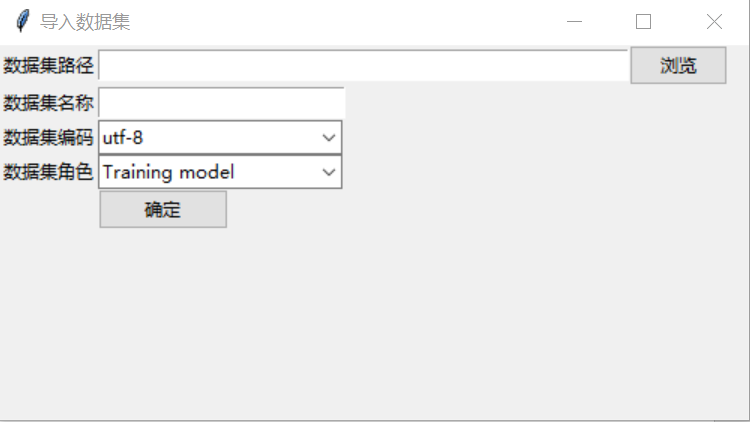
## 2.开始项目



有两种方式开始项目1.新建项目，2.导入已经存在项目。

新创建项目会产生一个project为后缀的文件，文件中有项目中模块信息。同时在该项目中创建的模块会保存在项目路径下。

## 3.导入数据集



导入的数据集文件为CSV文件，第一行为列明，编码可以为utf-8 ,gbk（强烈推荐用pandas output csv）数据集角色为['Training model', 'Reject', 'out of time sample','Score']

其中 'Training model' 为训练集，其他类型数据集不能参与模型训练。数据集变量可以分为['自变量', 'ID', 'TimeID', '目标', '以前模型分数'],其中自变量为输入变量，ID不参与分析过程，目标为模型的Target，以前模型模型分数是准备以后做分析预留的，TimeID会参与后面稳定性分析，建议把样本分为十个左右时间段。

## 4.数据集处理

4.1 分区

分区模块把数据集分为训练集和验证集（只针对'Training model'），分区方法有简单随机和分层抽样，其中分层抽样依据目标来进行分区。默认为无放回抽样，当训练集与验证集加和比例大于1时，则进行又放回抽样。如需设置分区后样本坏账率则把是否调整样本坏账率为‘是’，其中坏样本抽样比例为从总体样本中抽取多少百分比坏样本，整体样本坏账率为抽样后训练集和验证集坏账率



4.2 抽样

抽样与分区类似，其设置与分区基本一致，就不介绍了

## 5.交互式分组

交互式分组为单变量细分箱和粗分箱的过程，主要原理为1.数值型变量，首先对变量N等分（细分箱），然后对细分箱后变量进行粗分箱（应用了决策树算法），最后在进行人工的检查删选。2.字符型变量则依据坏账率排序然后用决策树进行粗分箱。3.对于有顺序的字符型变量（待开发）

5.1 数据集

交互式分组可以导入三组数据（训练集，验证集，拒绝集，时间外样本集），

5.2 预定义分组

变量设置：可以自己选择进行分组的变量（如果不进行分组则不能参与后面建模过程）

使用冻结分组：默认为否，如果为是，在该变量导入或者已经有分组信息的情况下不在重新分组，依照现有分组信息进行分组，如果没有则进行训练分组

导入分组信息：可以把其他的IGN后缀的分组信息进行导入

5.3 数值型变量分组设置

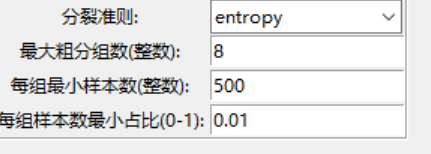
目前数值型只能进行分位数等分，可以输入整数细分箱个数，特殊值则是希望把该变量的某值作为单独一箱划分出来，不参与决策树分箱。其中缺失值单独一箱

5.4 字符型变量分组设置

可以选择数量或样本占比比较小的值不参与分箱直接归为others，也可以不限制

5.5 粗分箱设置

粗分箱设置则是对细分箱之后的箱子进行决策树分组，主要参数为



5.6变量拒绝设置

可以选择拒绝IV值小于某一阈值的变量参与后面建模流程

训练完成以后可以点击查看结果进入人工调整页面。双击变量名进入变量，双击细分组可以调整每个分箱的阈值，双击粗分组可以调整分组，在没有保存退出之前，图标需要点击刷新图标来重新显示 ，要不然只能显示刚开始的图标信息。

## 模型评分

6.1 入模数据

模型评分是建立评分卡的模块，需要导入数据分组模块。入模变量可以是分组后的WOE变量，或分组后整个变量的每个组，或每个免了的某一组，同IGN如果选择冻结变量则会固定现有模型的入模变量。

6.2 模型参数

可以选择模型参考标准，同时有三个变量筛选模式，forward,stepwise,no,以及P值设置，同时还有是否使用截距等。辅助变量和LASSO在变量数量比较多的情况下耗时会比较久，可酌情使用，以及部分评分卡设置。

6.3 模型训练完成以后可以查看报告以及导出报告

## 数据集打分

数据集打分可以给数据类型为Score的数据集打分