**生产环境部署python的必要性**

太平反欺诈模型是一个统计模型，涉及到数据分析计算。为适应现有的框架，模型被拆分为两部分，即现将统计计算的结果计算生成，并导入到数据库中，也即现在的模板；再根据新数据与模板的比对，来获取新数据的风险程度，也即相似度。这种方式将统计分析转化为了规则判断，在准确性上会有损失。

**以Python为中心的科学计算环境是现如今最主流的数据分析环境**，我们希望能在生产环境上部署以Python为中心的科学计算环境，主要是由于其带来的分析能力要远大于现有的java环境，**更稳定**、**更精确**、**更强大**。

**JAVA实现框架**

为配合java环境，将患者画像的聚合结果统计为模板，即存放于oracle的表格。统计结果会导致信息丢失，数据失去多维特征。

将数据分析以规则判定的方式实现，将患者画像所对应的群组简单表示为一系列规则。新增数据按oracle字段值的大小或种类来进行判别。

在现有框架之下，所有新增数据的计算必须经过java来实现，这意味着规则判定之外的处理方式均不可行。

在现有框架之下，难以实现有突破性的分析手段。

**Python 科学计算环境**

不存在模板的概念，患者画像在向量空间的聚合结果直接保存于向量空间中，为向量空间中的簇，后续直接使用该簇进行科学计算。

对于新增的数据，依旧使用向量空间中的患者画像来表示，并计算患者与聚合簇之间的距离、相似度。

由于使用向量空间来表示患者，患者画像的相似度计算可使用多种途径来计算。如：余弦相似度、欧几里德距离。

有更多可供选择的数据分析手段来帮助模型更好地鉴别异常值、检出案例。

基于统计分析的朴素贝叶斯分析等

基于机器学习的决策树、SVM等

基于深度学习的神经网络等

**JAVA实现框架**

源数据

（定期更新）

新增数据

（实时计算）

数据模板

相似度

计算结果

优点：

1. 现有框架可用

缺点：

1. 过程复杂，更新模型牵连人员广，沟通复杂，迭代慢，维护不便。

2. 基于java的规则判断框架，数据分析手段少，分析准确性欠佳。

**JAVA下的数据流**

虚拟机

（二楼）

UAT环境

核心

生产环境

**Python 科学计算环境**

源数据

（定期更新）

新增数据

（实时计算）

Python

科学

计算

环境

相似度

计算结果

优点：

1. 科学计算环境稳定，模型结构清晰，数据流输入输出明确，中间过程可控，维护简单，新需求迭代快。

2. 数据分析手段增多，功能强大，准确性高，可探索性强，能更好地服务于具体业务。

缺点：

1. 新的科学计算环境，需要客户配合搭建。

**Python 下的数据流**

核心

生产环境

优化：

在生产环境上，直接用生产数据生成稳定模型和向量空间，在本地存储。后续直接调用模型来判断新增案件的风险水平。

省掉了与虚拟机有关的中间流转，使流转过程更简明、可靠、易维护，不易出错。

**太平现场部署Python环境的步骤**

在太平现场部署Python环境的环境要求：

- Oracle客户端或者是Oracle instant client客户端

- 在Linux系统下新建普通帐户（例TPFQZ），专门用于反欺诈模型部署。

在帐户TPFQZ下，需要新增的软件环境为：

- Python科学计算发行版本Anaconda，版本号4.4.0（推荐）以上。

- 在该帐户下的Python环境中安装用于连接Oracle的模块cx\_Oracle

**Anaconda安装步骤：**

1. 前往清华大学开源镜像站下载Anaconda安装文件:

https://mirrors.tuna.tsinghua.edu.cn/anaconda/archive/

推荐下载:

Anaconda3-4.4.0-Linux-x86\_64.sh

2. 将该文件拷贝入帐户TPFQZ目录下，运行:

bash Anaconda3-4.4.0-Linux-x86\_64.sh

跟据向导完成安装。

3. 将Python 路径添加入.bashrc中。例如:

export PATH="/home/TPFQZ/software/anaconda3/bin:$PATH"

**安装 cx\_Oracle 模块：**

1. 前往pypi下载Oracle官方用于Python 连接Oracle 的模块：

https://pypi.python.org/pypi/cx\_Oracle

下载对应Python3和64位的版本：

cx\_Oracle-6.1-cp36-cp36m-manylinux1\_x86\_64.whl (md5)

2. 将该文件拷贝入帐户TPFQZ目录下，运行：

pip install cx\_Oracle-6.1-cp36-cp36m-manylinux1\_x86\_64.whl

3. 配置Oracle 环境变量，将相关设置加入.bashrc中：

export ORACLE\_HOME=Oracle或Oracle instance client 地址

export PATH=$ORACLE\_HOME:$PATH

export LD\_LIBRARY\_PATH=$ORACLE\_HOME:$LD\_LIBRARY\_PATH

export NLS\_LANG='simplified chinese\_china.ZHS16GBK'

**环境应用风险**

因为要在生产环境上直接部署数据模型，故可能存在相关风险，列举如下：

1. 因需要在生产环境上创建单独的帐户来部署反斯诈系统，故该帐户存在影响系统正常工作的风险，如修改相关的系统文件和参数。

规避措施：该帐户对外只开放访问Oracle的权限。

2. 数据模型在运算的时候，需要占用CPU、内存等资源，所以服务器有宕机的风险。

规避措施：对该帐户设置资源分配上限，建设为总量的70%。

3. 模型需要读取业务表的权限，以及更新T\_SGZSXJK和T\_DBZZYYC两张业务表的权限。所以使用这两张业务表的服务有报错的风险。

规避措施：规范代码的行为，作好上线前的测试。