# 需求

借鉴美国模型团队经验,优化反欺诈模型。本次美国模型培训内容分以下7部分, 具体内容请参见附件PPT。

Session 1: Introduction to Fraud Modeling in SAS FM

Session 2: Message Layouts in SAS FM

Session 3: Signatures

Session 4: Data Preparation

Session 5: Model Development

Session 6: Performance Evaluations

Session 7: Model Packaging

经过比较与之前做法存在以下不同:

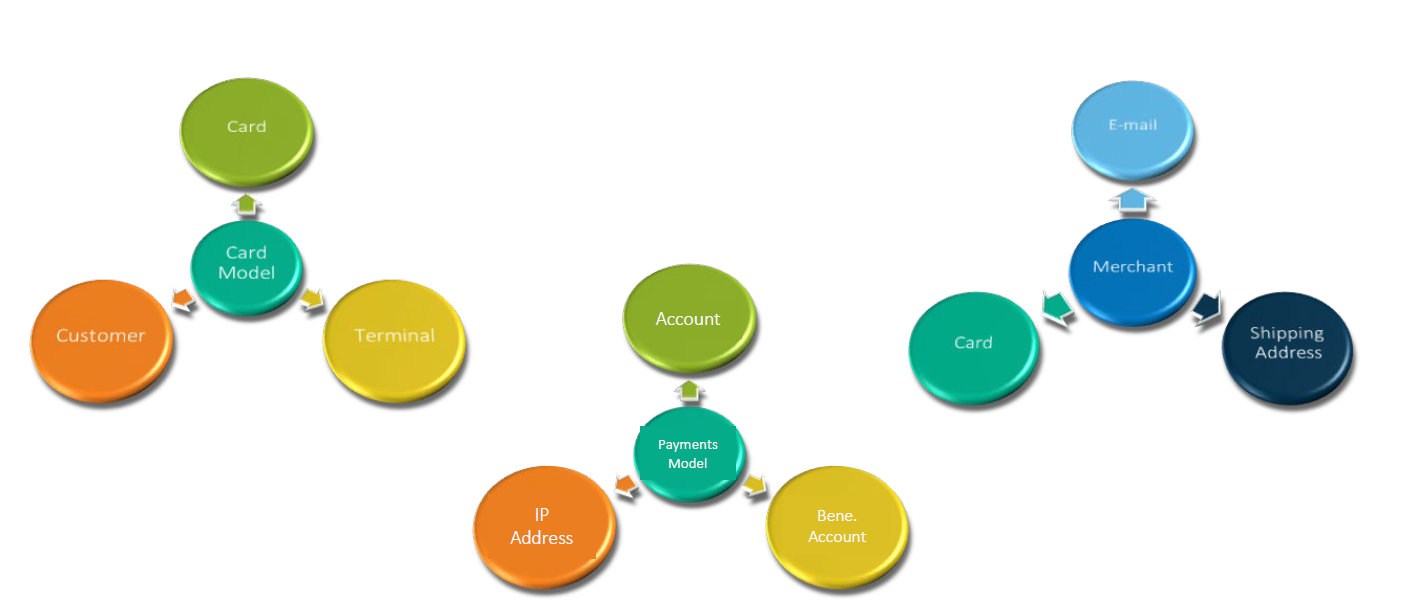
* **数据源名字统一处理**

开发模型之前需要将TDR不同数据源名字统一处理,如交易金额在CCCA表中为tca\_client\_amt 而在CCBF中为 tbt\_tran\_amt。

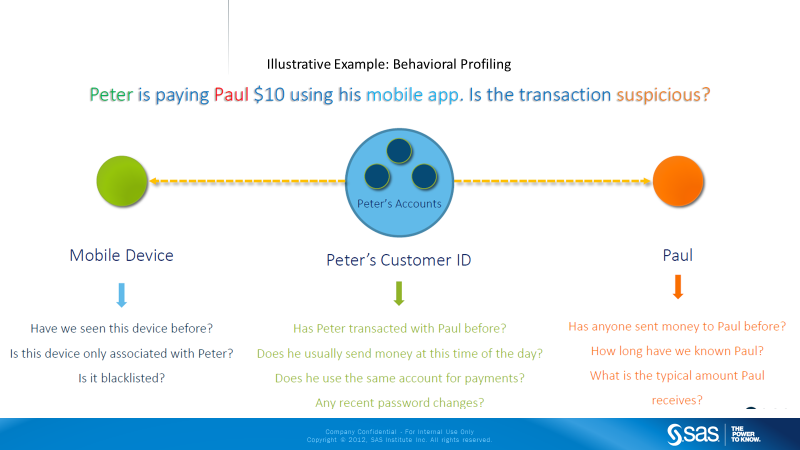
* **模型特征维度**

构建模型特征会有2-3个Level的特征维度，

如下图卡模型涉及卡号/商户号/客户号3个维度的特征:



如下图转账模型及设备/客户号/收款人3个不同维度:



名词解释:

特征 -为模型的输入变量,如最近1小时交易金额。

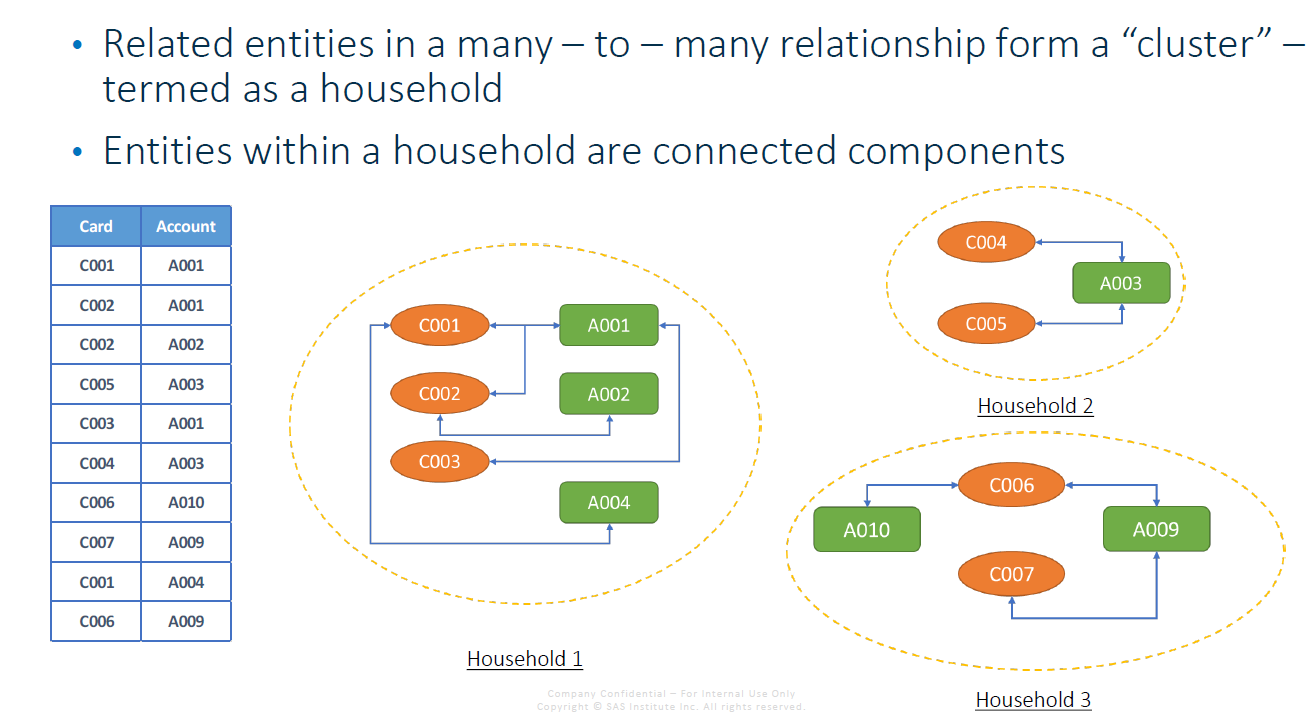
Signatures-为构造特征变量在MEH中存储的中间信息，如为了计算” 最近1小时交易金额”需要在MEH存储最近100笔交易的交易金额，该100个交易字段为Signatures。

* **扩充模型特征**

美国团队候选模型特征为30000个左右，尝试扩充模型特征。

* **抽取样本**

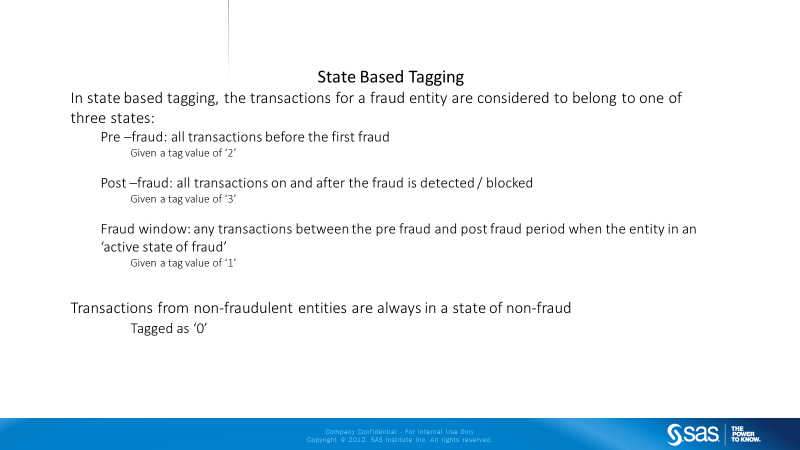
美国团队采取Household级别进行抽样，该方法可以保证抽样偏差较小。



* **欺诈标签**

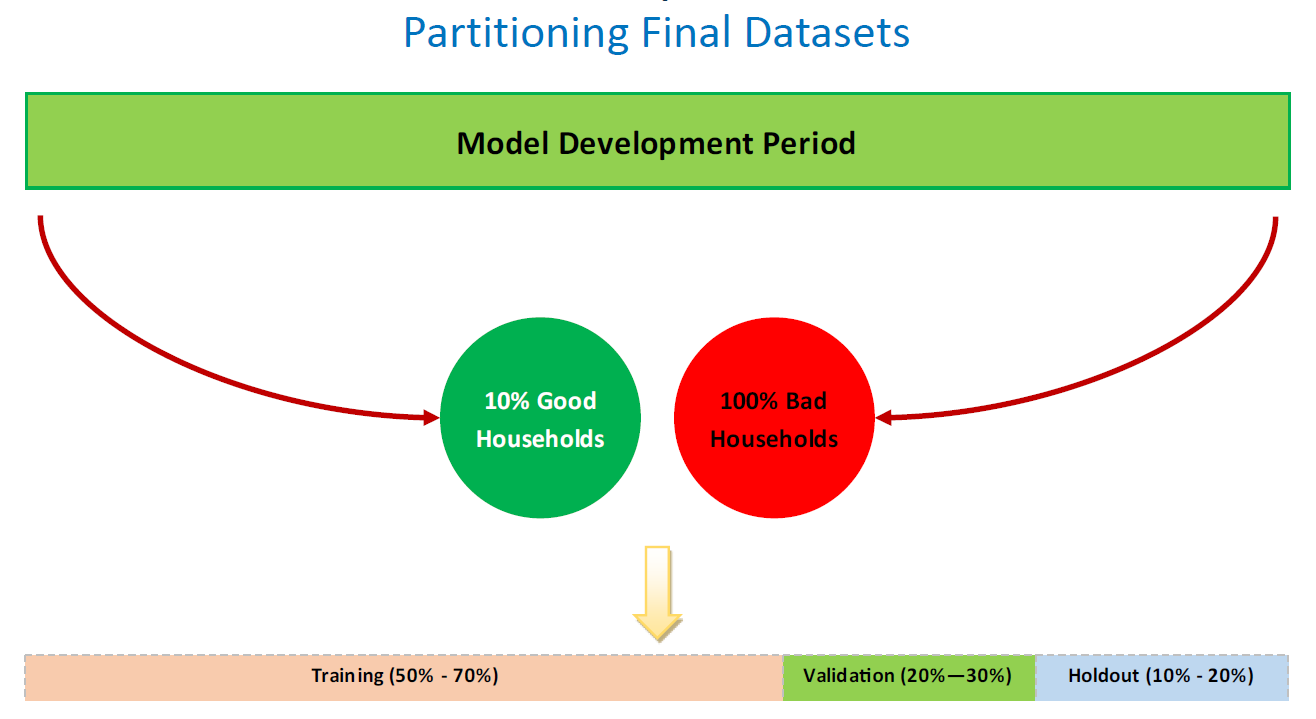
美国团队打标签也会采取模糊匹配方式，如欺诈时间的上下浮动。

训练模型有2中不同Target,一种为状态级别,另一种为交易级别，2中级别各有优缺点，在实际运用中状态级别较多。



* **数据分割**

在分割Training/Validation/Holdout数据时需要考虑时间维度。



* **逐月检查原始数据的分布**
* **搭建BaseSAS SFM**

美国团队开发模型环境均由BaseSAS搭建而成,用于模拟生产SFM,其中涉及Error checks/Scoring exclusions/Replacement window/Tag based filtering/Load essential reference datasets as hashes

* **分群**

分群方法有Hard vs soft segmentation,我们之前均基于业务建议的分群方法，后续可尝试基于数据的分群方法。

* **RiskTable**

美国团队构建模型时使用RiskTable较多，可能会涉及200-300个。

* **模型评估**

与之前评估方法类似，但是在评估时添加了FraudWindow的概念。

# 数据

这次为优化模型非首次开发模型，所以直接提取SFM TDR数据可保证开发数据与生产数据格式及内容保持一致。

数据源 :SFM TDR数据

数据时间窗口:近1年半

# 方案

* **数据源名字统一处理**

在模型中添加预处理模块。

* **模型特征维度**

可选维度:Account number/Customer ID/Card number/Beneficiary account number

/Beneficiary bank/Merchant ID/Terminal ID/IP address/Shipping address。

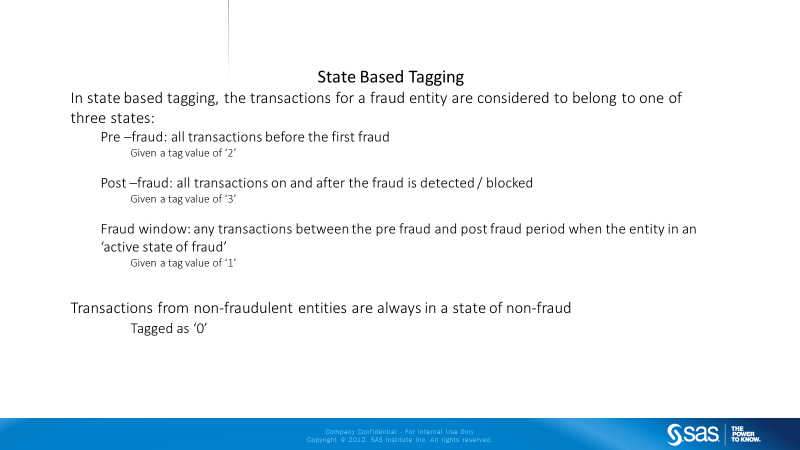
* **扩充模型特征**

增加模型特征，运用GA算法补充特征变量。

* **抽取样本**

用Household方法抽样。

* **欺诈标签**

****

* **数据分割**

在分割Training/Validation/Holdout数据时需要考虑时间维度。

* **逐月检查原始数据的分布**
* **搭建BaseSAS SFM**

搭建BaseSAS SFM,便于模型快速开发部署。

* **分群**

尝试基于数据的分群方法。

* **RiskTable**

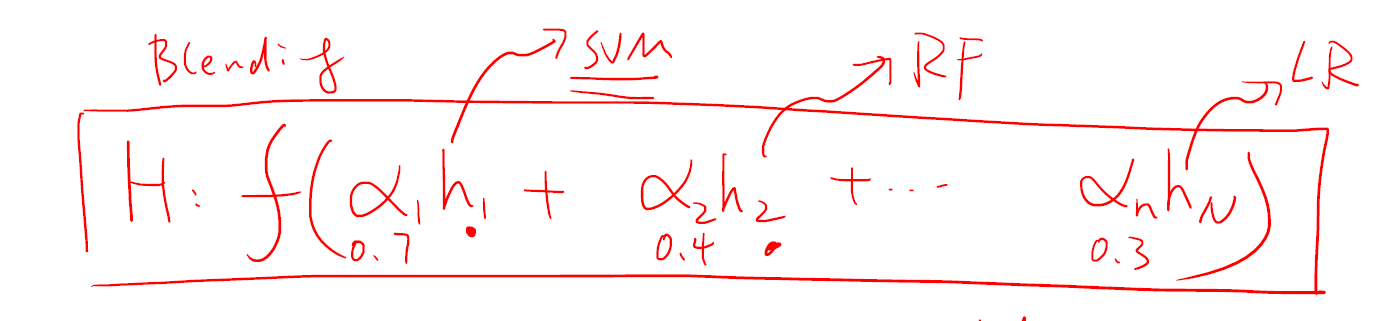
扩充RiskTable。

* **模型评估**

增加FraudWindow概念。

* **模型训练**

增加模型融合步骤



* **部署模型**

在SFM测试环境部署模型，计算评分，确保和模型开发阶段的评分一致

# 计划

以下是我的初步计划，具体看到数据后会进行调整。

|  |  |  |  |
| --- | --- | --- | --- |
| **Steps** | **Days** | **Start** | **End** |
| **DataPrepare** | **8** |  |  |
| Fraud tagging | 3 |  |  |
| Sampling | 3 |  |  |
| Analyze\_data | 2 |  |  |
|  |  |  |  |
| **Modeling** | **30.5** |  |  |
| ErrorChecks/Exclude | 1 |  |  |
| Segmentation | 3 |  |  |
| Risk tables | 5 |  |  |
| Feature generation | 10 |  |  |
| Feature selection | 5 |  |  |
| Model training | 5 |  |  |
| Ensemble | 1 |  |  |
| Score Alignment | 0.5 |  |  |
|  |  |  |  |
| **Performance&Packaging** | **1** |  |  |
| Performance Evaluations | 0.5 |  |  |
| Model Packaging | 0.5 |  |  |
|  |  |  |  |
| **Total** | **39.5** |  |  |

# 目标

如上述方案中所描述,本次优化模型过程中增加以下内容:

1.搭建BaseSAS SFM

2.增加数据源名字统一处理模块

3.增加模型特征维度

4.扩充模型特征

5.调整抽取样本方法

6.调整欺诈标签方法

7.调整数据分割方法

8.增加逐月检查原始数据的分布功能

9. 调整分群

10.增加RiskTable

11.调整模型评估方法

12.调整模型训练方法增加模型融合步骤

13.在SFM测试环境部署模型，计算评分，确保和模型开发阶段的评分一致

# 附录

A Primer on Model Development and Deployment: