PPD风控算法设计大赛

团队：数据匠

Email: chensheng0802@outlook.com

任务分析

- 总体任务：设计风控算法模型，预测1万新用户6个月内发生逾期的概率。
- 已知条件：

(1).8万用户历史数据（训练集）

a. 个人信息

b. 行为信息

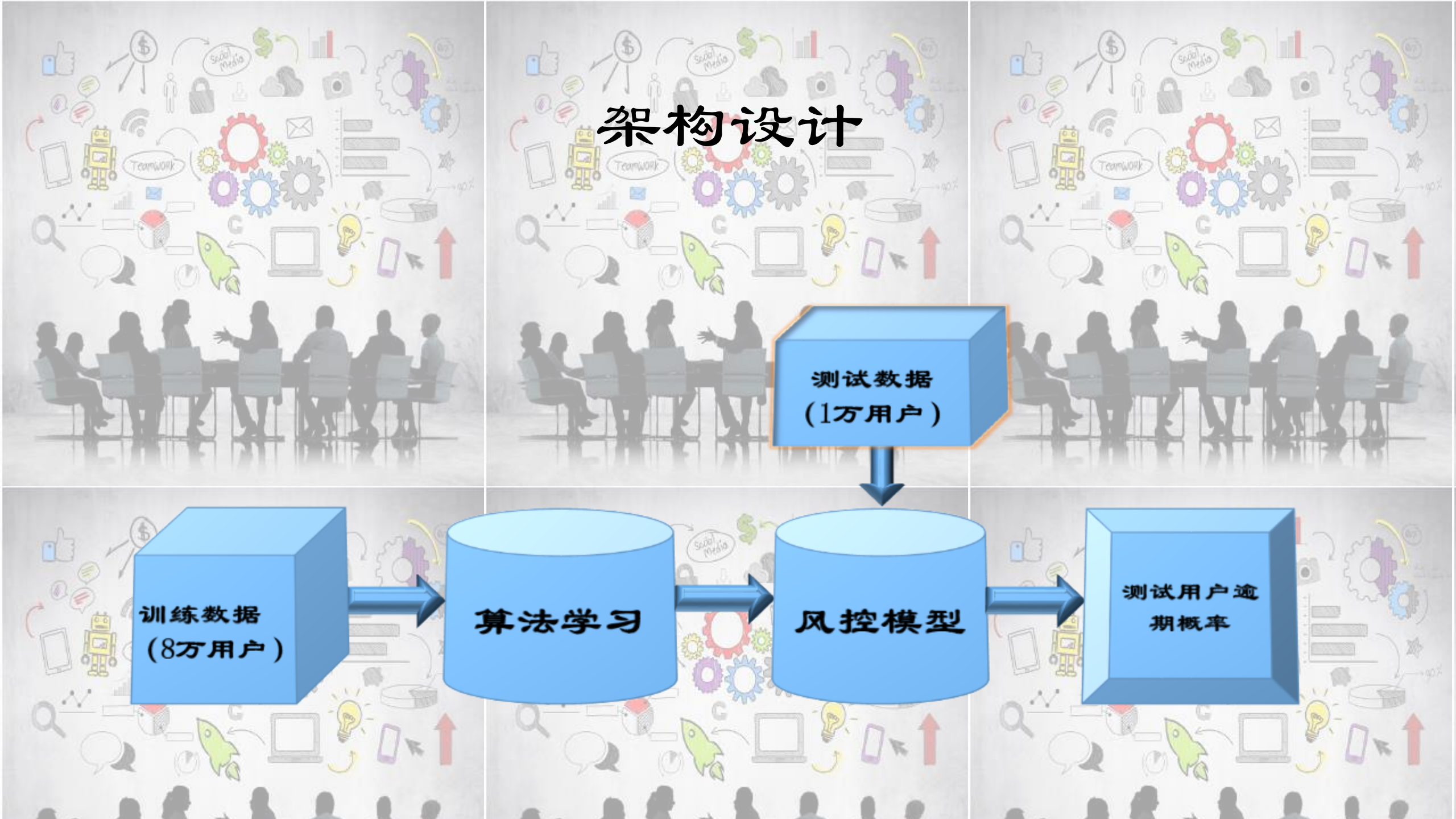
c. 历史逾期标签

(2).1万用户数据（测试集）

a. 个人信息

b. 行为信息

架构设计



总体方案

- ✓数据初步分析
- ✓数据清洗
- ✓特征工程
- ✓模型训练
- ✓模型评估
- ✓预测结果



数据初步分析

数据集总共分为3份：

1. **Master Data(209维):** 主数据集

a. UserInfo(24维)： Categorical(22维) + Numerical(2维)

b. Education_Info(8维)： Categorical(8维)

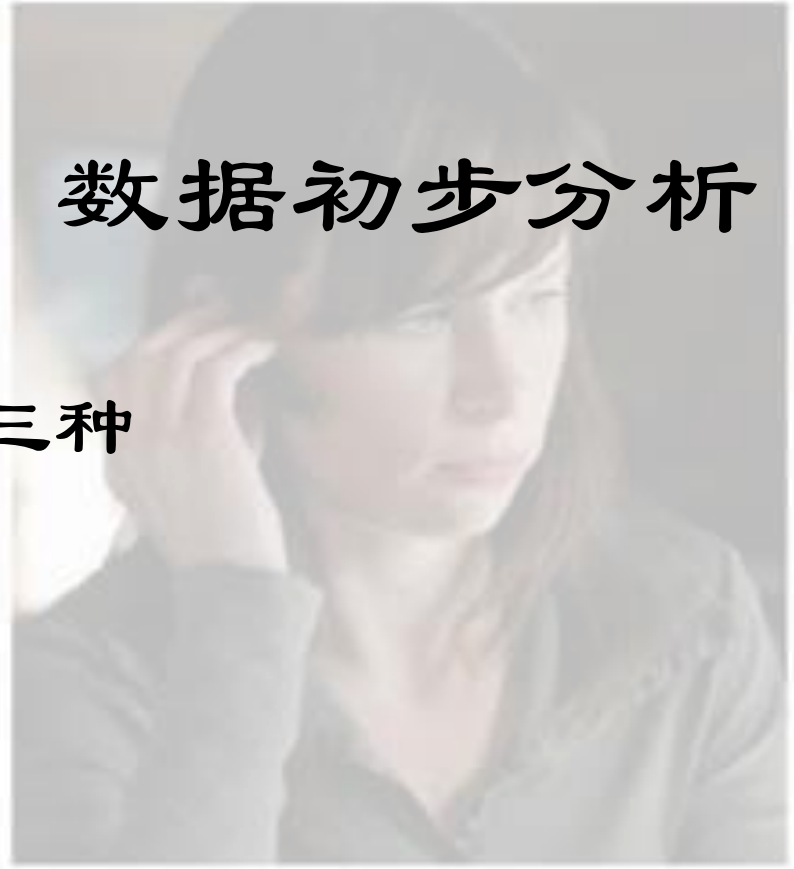
c. WeblogInfo(58维)： Categorical(3维) + Numerical(55维)

d. ThirdParty_Info(119维)： Numerical(119维)

2. **UpdateInfo Data(3维):** 用户更新个人信息记录

3. **LogInfo Data(4维):** 用户操作行为日志

数据初步分析



- 数据类型主要有三种
- 1. 类别型
- 2. 连续型
- 3. 缺失值

特征工程

数据和特征决定机器学习的上限，模型和算法则逼近这个上限。

Question:

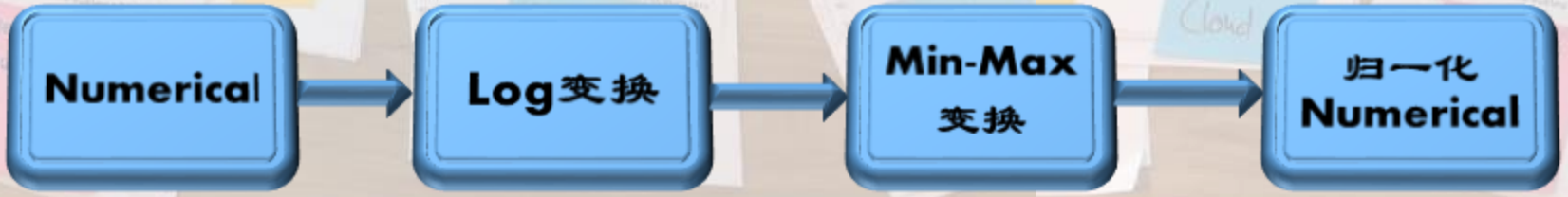
1. 如何编码才能使特征适合模型的学习？
2. 如何编码才能使信息不失真？

思路决定出路！

特征工程<设计思路一>

1.类别型变量

2.连续型变量

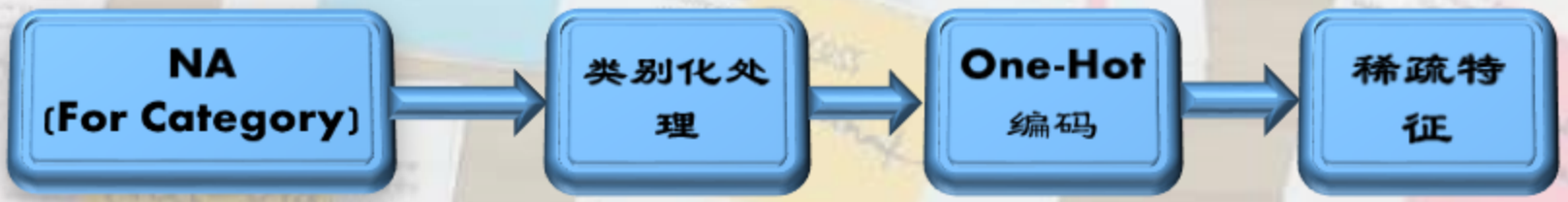


特征工程<设计思路二>

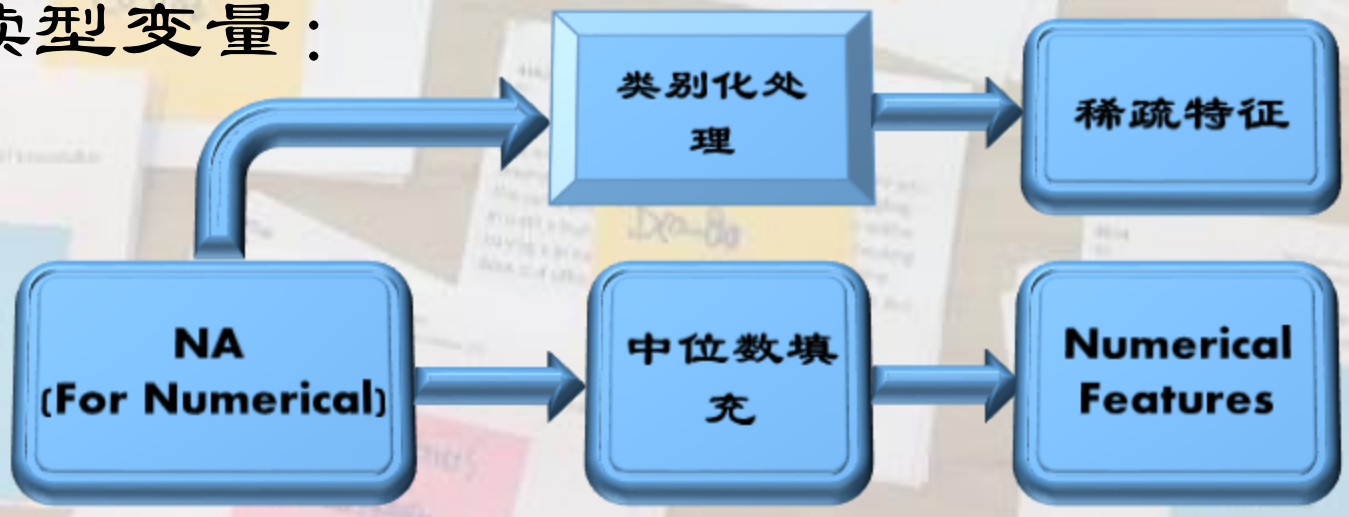
3.缺失值

NA is also the Voice of the Customers.

(1).类别型变量：



(2).连续型变量：



模型设计

初赛：

1.XgBoost

2.GBDT

复赛：

1.XgBoost

2. 非均等代价：

a. 给逾期记录设置更大代价进行学习

b. 代价比=二类记录出现频数反比

模型评测

1.线下AUC(CV = 5): AUC = 0.769

2.线上测评: AUC = 0.7703

线上线下对比，模型较为稳定

后续思考

- 1. 通过预测结果和真实结果对比，可以看出模型(AUC=0.77)倾向于将用户判别为非违约，这与模型设计的初衷是相反的；
- 2.是否是个模糊识别的问题？表现在：

a. 有用户行为正常，违约

b. 有用户行为异常，但不违约

- 3.改进点：是否可以将二分类转化成多分类任务，实现更加精准的用户划分？比如按违约级别分类

a. 通过模型1将用户划分成不同违约级别的种类；

b. 通过模型2预测用户属于哪个违约级别；

Thank for your attention!