多源医生主数据匹配项目报告

1. 项目背景
2. 算法设计
3. 匹配分析结果
4. 优化建议

Author：罗红

项目背景

公司需要将内部不同来源的医生数据（CRM系统、市场活动数据）与Veeva提供的医生主数据进行统一匹配。

- 目标是构建**黄金客户视图（Golden Customer View）**，解决数据重复和不一致问题。
- 数据来源：

1.veeva_master_doctors.csv（100条记录）

- doctor_id, name, hospital, department

2.customer_A_doctors.csv（45条记录）

- internal_id, doctor_name, work_unit, dept

3.customer_B_doctors.csv（55条记录）

- id, physician_name, hospital_name, specialty

算法设计

1.数据预处理

- **去空格、统一大小写**：将所有文本字段统一为小写，去除首尾空格。
- **全角转半角**：避免中文全角字符导致匹配失败。
- **字段标准化**：
  - cust_a: doctor_name → name_norm，work_unit → hospital_norm，dept → specialty_norm
  - cust_b: physician_name → name_norm，hospital_name → hospital_norm，specialty → specialty_norm
  - master: name → name_norm，hospital → hospital_norm，department → specialty_norm

2.关键匹配字段

- **姓名（name）**：最核心字段，区分医生身份
- 医院（hospital）：辅助验证，防止同名医生匹配错误
- 科室（specialty/department）：进一步约束，提高匹配精度

3.匹配逻辑

1. **姓名相似度**：
   - FuzzyWuzzy token_sort_ratio（处理字符顺序差异）
   - Jaro-Winkler 距离（处理小拼写差异）
   - 取两者平均值作为姓名综合得分
2. **医院相似度**：
   - 精确匹配得 1.0
   - 子字符串匹配得 0.8
   - 模糊匹配（partial_ratio）取分数
3. **科室相似度**：
   - 精确匹配得 1.0
   - 子字符串匹配得 0.8
   - 模糊匹配（partial_ratio）取分数

4.综合得分与匹配状态

1. **加权组合**：
   - score = 0.6 * name_score + 0.25 * hospital_score + 0.15 * specialty_score
2. **匹配状态判定**：
   - match：score ≥ 0.75
   - possible_match：0.6 ≤ score < 0.75
   - no_match：score < 0.6

阈值设定流程

1.定义目标：
区分veeva主数据中的医生和客户数据表中的医生是否是同一个人

2.构建匹配得分

将医生姓名、医院名称、科室名称相似度加权

3.设定候选阈值区间

- 遍历 0.50 ~ 0.90（每隔 0.05）
- 计算每个阈值下的匹配率 / 成功率

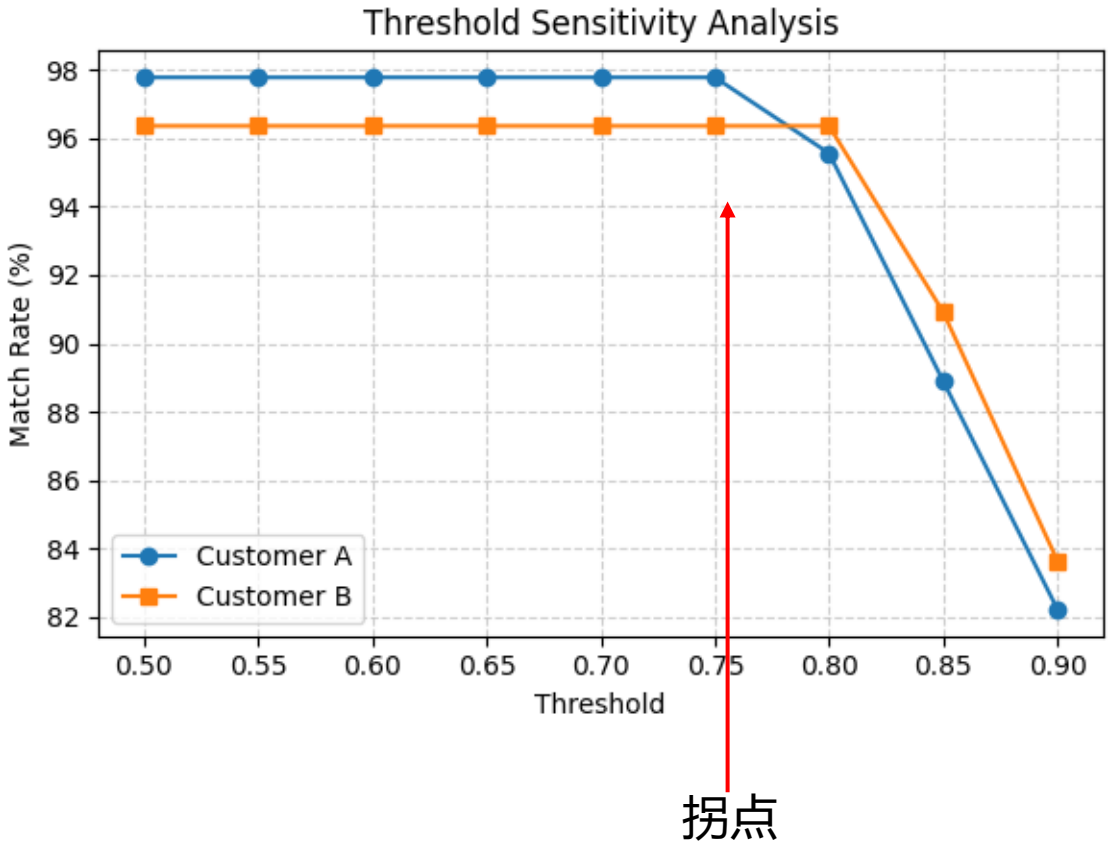
4.阈值敏感性分析

生成分析表格（Threshold vs Match Rate）绘制曲线图，观察拐点

5.选择合理阈值

在 准确性 与 覆盖率 之间平衡选取 **0.75** 作为最终阈值：

- 匹配成功率较高
- 明显减少误匹配（保证质量）



匹配分析结果

模型匹配结果					
客户	总记录数	成功匹配	可能匹配	未匹配	成功匹配率
客户A	45	44	1	1	97.78%
客户B	55	53	0	2	97.78%

未（完全）匹配情况分析				
客户	医生名称	医院	科室	可能原因
A	孙悦	安贞医院	心内科	Veeva主数据中不存在
A	李娜娜	北大三院	血液科	客户名单中医生数据为李娜娜，Veeva名单中为李娜，医院和科室一致，可能是**名字登记不够准确**
B	周涛	复兴医院	月坛社区	Veeva主数据中不存在
B	吴刚	武警总医院	骨科	Veeva主数据中不存在

策略优化建议

1.医院和科室标准化字典:

1. 维护常见医院别名（如“301医院” → “北京301医院”）
2. 科室同义词映射，提高模糊匹配准确度

2.多阶段匹配:

1. 第一轮严格匹配（高阈值）
2. 第二轮宽松匹配（低阈值）并人工复核

3.引入机器学习方法:

1. 使用特征（姓名相似度、医院相似度、科室相似度）训练二分类模型
2. 自动预测是否为同一医生，可提高复杂情况下匹配率

4. 动态更新权重:

1. 基于业务开展频次，月度或者季度根据客户或者运营人员反馈，动态更新特征权重