数据挖掘导论lab3报告

3220106424 张天逸

数据集概况

Kaggle上的Bank Marketing Dataset这是经典的市场营销银行数据集，最初上传在UCI机器学习存储库中。该数据集提供了关于金融机构市场营销活动的信息。

数据特征

特征名称	数据类型	物理意义	取值范围 / 取值列表
age	连续值	客户年龄	18–95（整数，岁）
job	类别型	职业类型	`admin.`, `technician`, `services`, `management`, `retired`, `housemaid`, `unemployed`, `student`, `entrepreneur`, `blue-collar`, `self-employed`, `unknown`
marital	类别型	婚姻状况	`married`, `single`, `divorced`
education	类别型	教育水平	`primary`, `secondary`, `tertiary`, `unknown`
default	类别型	是否有信用违约记录	`yes`, `no`
balance	连续值	账户余额	-6847–81204（欧元，整数）
housing	类别型	是否有房贷	`yes`, `no`
loan	类别型	是否有个人贷款	`yes`, `no`

特征名称	数据类型	物理意义	取值范围 / 取值列表
contact	类别型	联系方式	`unknown`, `telephone`, `cellular`
day	离散值	最后一次联系的日子	1–31（整数，日）
month	类别型	最后一次联系的月份	`jan`, `feb`, ..., `dec`
duration	连续值	最后一次联系通话时长	0–5 000+（秒，整数）
campaign	离散值	本轮营销对该客户的联系次数	1–50+（整数）
pdays	离散值	上一次营销距今天数；-1表示未曾联系	-1, 0–500+（整数）
previous	离散值	之前营销中对该客户的总联系次数	0–20+（整数）
poutcome	类别型	之前营销的结果	`unknown`, `other`, `failure`, `success`
deposit	类别型	客户是否订阅定期存款	`yes`, `no`

数据分布（以是否订阅为例）

类别	样本数量	占比
未订阅 (no)	5 873	51.9%
已订阅 (yes)	5 289	48.1%

> *注：总样本数 11 162 条，Imbalance Ratio ≈ 1.11（多数/少数）。*

部署

依赖：pandas numpy scikit-learn matplotlib scipy mlxtend

如果存在python版本冲突，建议创建虚拟环境：

cd lab3 python -m venv .venv # 或 conda create -n pcos python=3.12 source .venv/bin/activate # Windows: .venv\Scripts\activate

算法

Apriori

1. 事务集输入 将每个客户看作一个事务（transaction），事务内由“特征=取值”形式的项组成。
2. 一项集计数 遍历所有事务，统计每个单项（1-项集）的出现次数。
3. 频繁1-项集筛选 根据最小支持度阈值，去掉出现次数不足的单项，得到频繁1-项集。
4. 候选k-项集生成（连接与剪枝）

从频繁(k–1)-项集中，两两做“并集”，仅保留大小正好为k的组合。

对每个候选项集，若它的所有 (k–1) 子集都出现在频繁(k–1)-项集中，则继续，否则剪枝。

5. 候选项集计数与筛选 遍历所有事务，对每个候选k-项集检查其是否为事务的子集，并累加计数；随后根据支持度阈值筛选出频繁k-项集。
6. 迭代直至无新频繁项集 重复第4、5步，k从2开始递增，直到生成不了新的频繁项集为止。
7. 结果整合 将所有频繁项集合并输出，支持度可由计数除以事务总数得到。

FP-Growth

1. 构建头指针表 统计所有单项的全局出现次数，筛出满足最小支持度的频繁单项，并按降序记录在头指针表（header table）中。
2. 构建FP-Tree

建一个根节点（空）。

对每个事务：过滤并保留频繁单项，按全局频次降序排列，从根节点沿分支插入/更新各项节点，同时维护头指针表中的“节点链”以便快速跳转。

3. 递归挖掘条件FP-Tree 对头指针表里最不频繁的单项开始：

前缀模式基：沿着该项在树中的所有节点链，逆向抽取其到根节点的路径及节点计数，

以这些路径重新构建一个小规模的“条件FP-Tree”，

在条件树上重复上述头指针表→FP-Tree→递归挖掘过程，生成以该项为后缀的所有频繁模式。

4. 无候选集显式生成 FP-Growth无需像Apriori那样生成大量候选组合，通过树结构和头链完成频繁项集的高效统计。

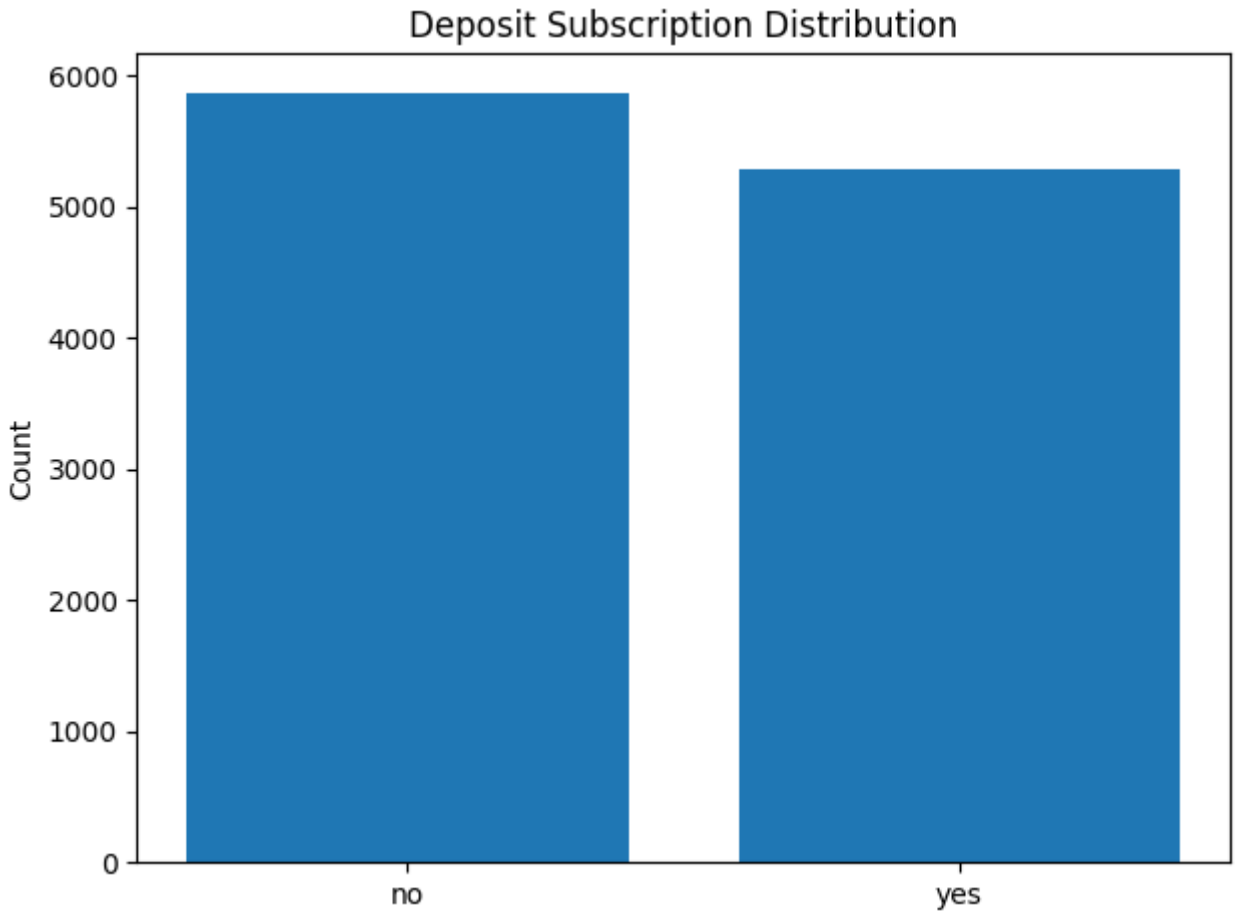
数据预处理

1. **缺失值检查**

```
# 检查是否存在缺失的列
missing_cols = [col for col in categorical_features if col not in df.columns]
if missing_cols:
    print(f"Missing columns: {missing_cols}")
    exit()
else:
    # 将数据转换成适合关联规则挖掘的0-1矩阵
    df_encoded = pd.get_dummies(df[categorical_features])
```

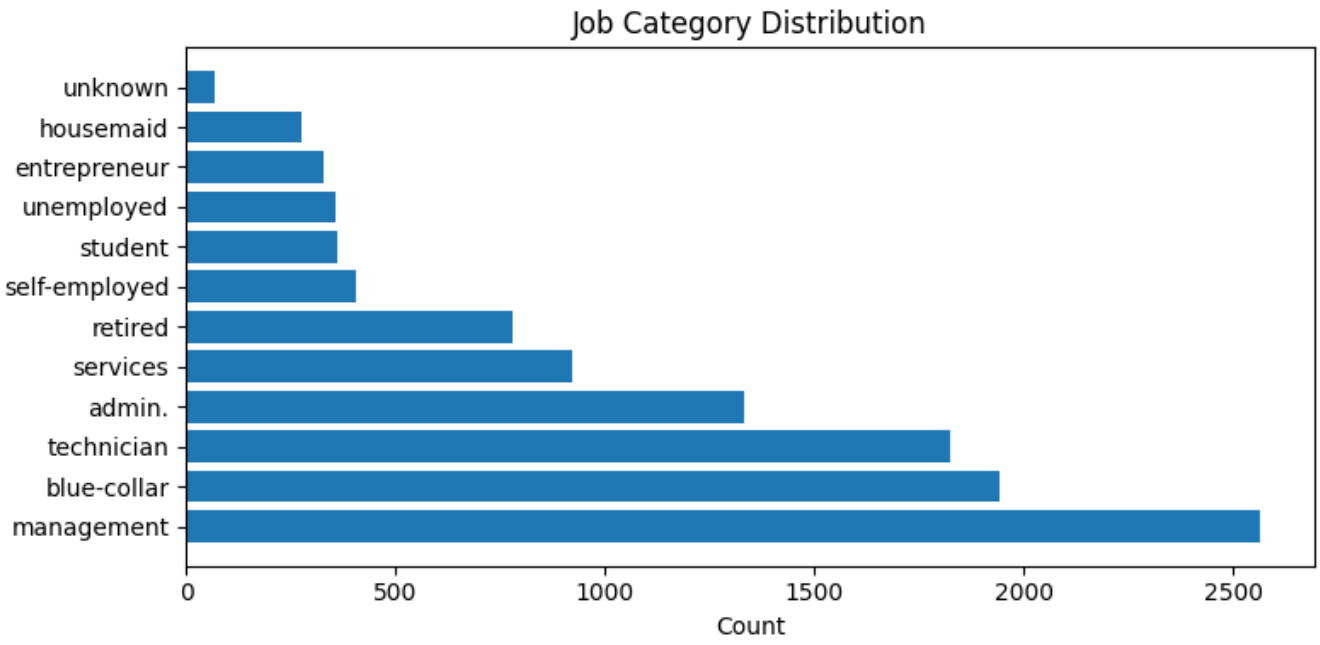
可视化选型及结果

Deposit 订阅分布（柱状图）



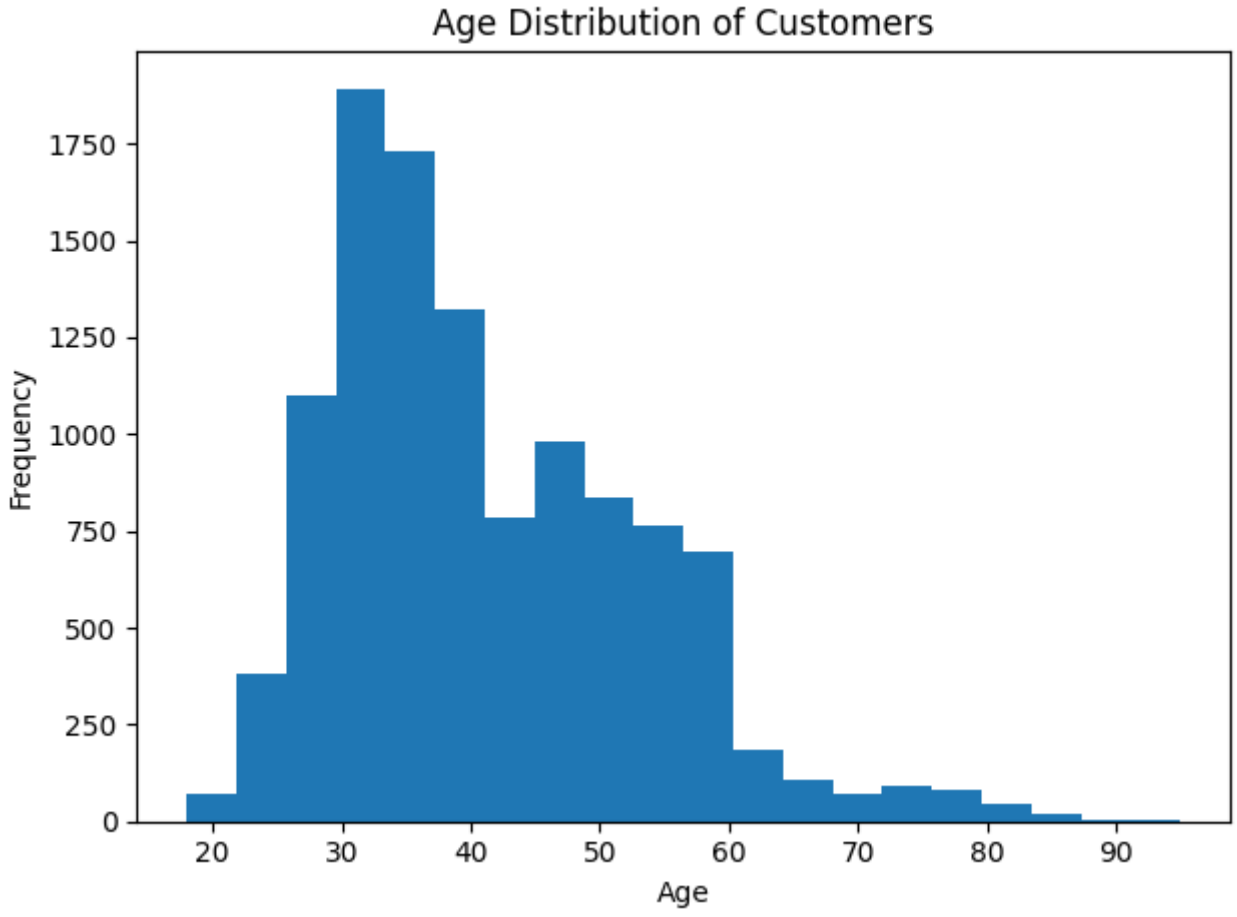
> - 说明数据较平衡

职业分布



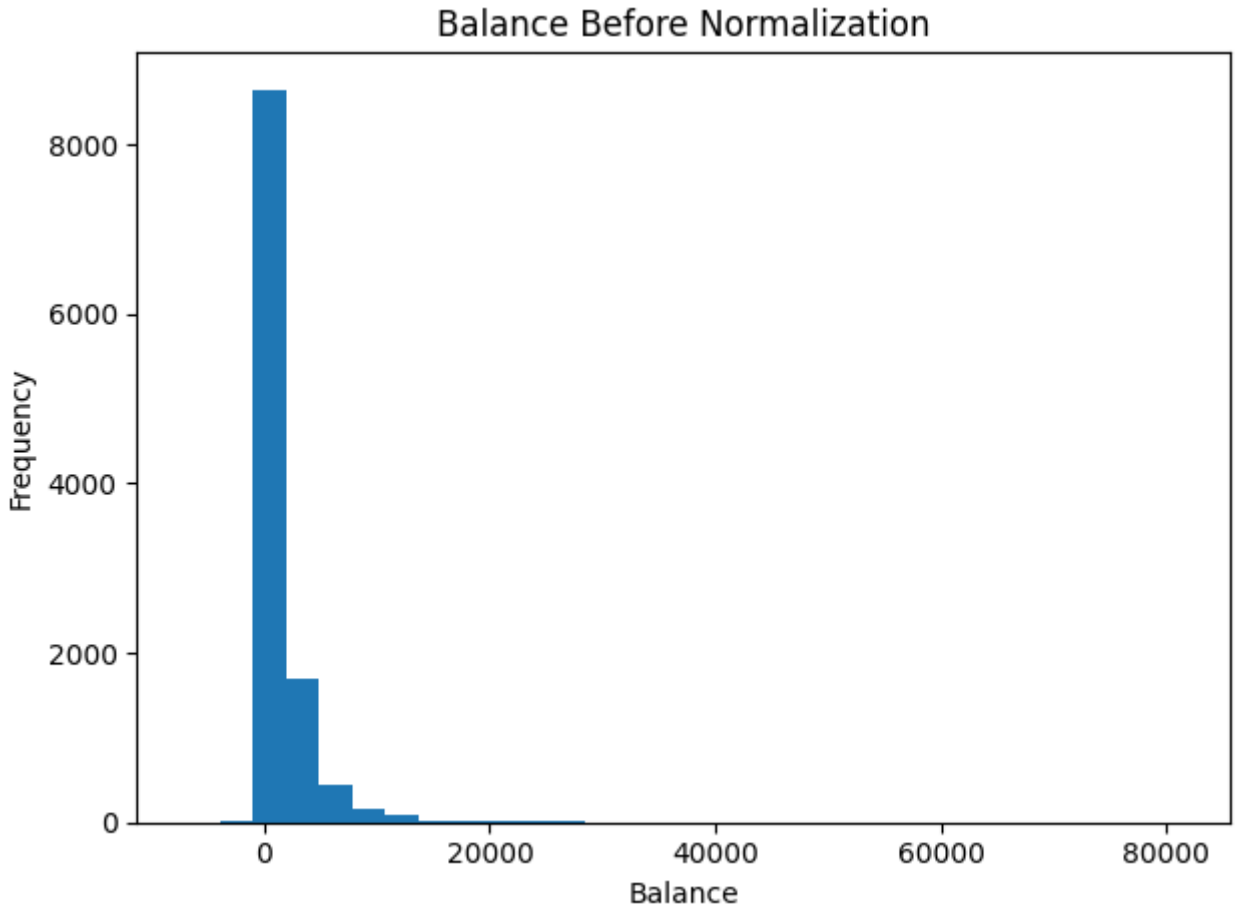
- `management`、`blue-collar`、`technician` 为主力职业；
- 少数职业如 `housemaid`、`entrepreneur` 样本较少，可视为"长尾"类别。

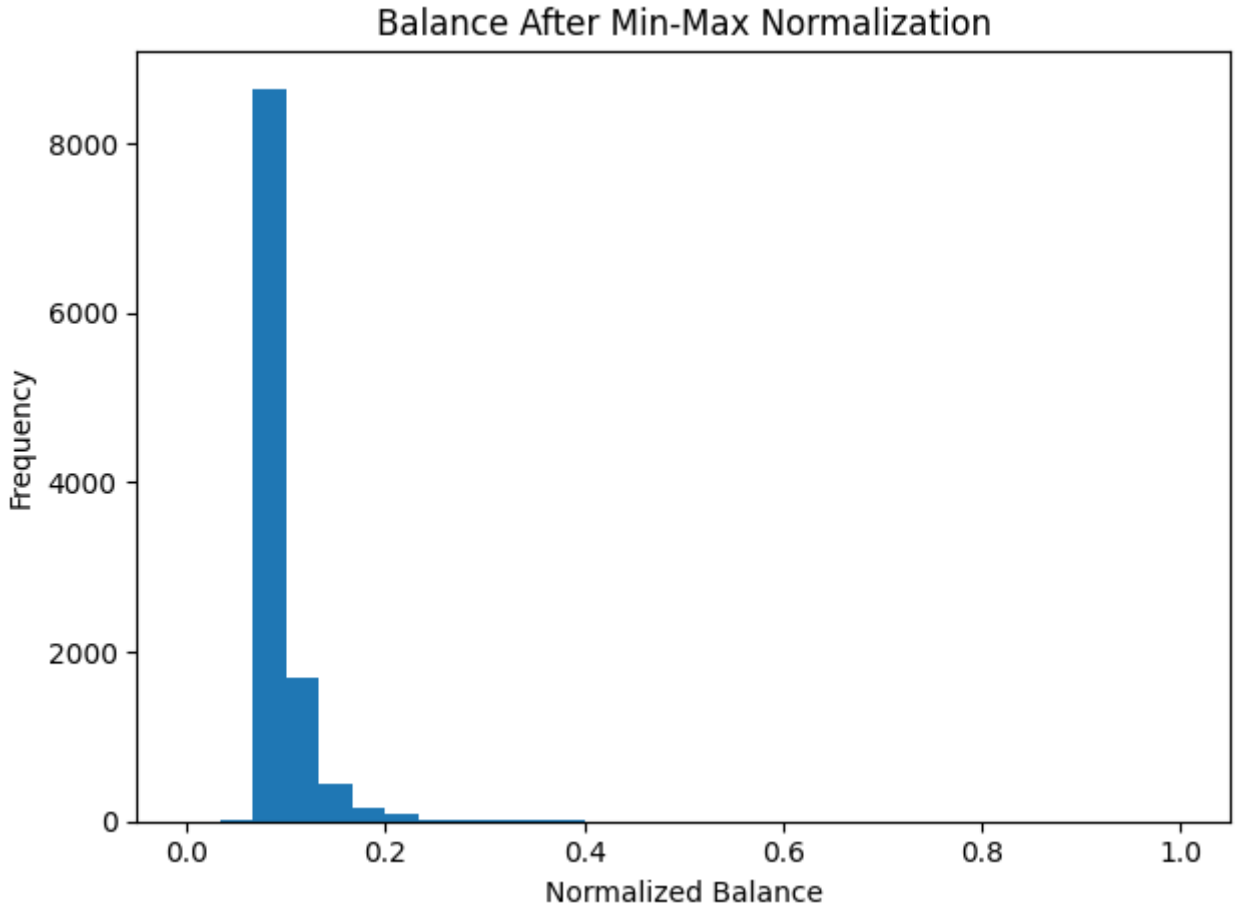
年龄分布



- 30-60 为主要年龄；

归一化前后 Balance 分布对比（双图）


Balance Before Normalization
Frequency
8000
6000
4000
2000
0
0
20000
40000
60000
80000
Balance


Balance After Min-Max Normalization
Frequency
8000
6000
4000
2000
0
0.0
0.2
0.4
0.6
0.8
1.0
Normalized Balance

> **结果说明**:

> - 归一化后，`balance_norm` 均匀映射到 `[0,1]` 区间，便于不同量级特征的可视化比对或机器学习输入。

四、结果分析与讨论

Apriori 与 FP-Growth 的频繁模式对比

- **一致性验证**：两种算法给出的 Top10 频繁项集完全相同，说明手写实现正确且可靠。
- **模式含义**：
  - 最高频的是 `default=no`（98.5%），几乎所有客户都无违约记录。
  - `loan=no`（86.9%）、`default=no ∧ loan=no`（85.9%）等，均反映了大多数客户既无信贷违约，也未申请个人贷款的"常态"特征。
  - `poutcome=unknown`（74.6%）说明绝大多数客户在本次营销前未曾联系过。

关联规则 Top-5 解读

```
Top 5 Association Rules:
{poutcome=unknown, deposit=no, month=may}
  → {housing=yes, contact=unknown, default=no, marital=married}
    support=0.057,
    confidence=0.403,
    lift=4.603,
    leverage=0.045,
    chi2=2298.4 (p≈0.000)
…
```

- **方向**：所有这些规则的 **后件** 都是 "deposit=no" 以及与之高度相关的特征组合，说明挖掘出的最显著模式主要是**"不订阅"** 客户的特征。
- **支持度 ~5.7%**：每条规则只覆盖了约 6% 的记录（一个较小的子集），意味着模式虽显著，但并不代表整表多数客户。
- **置信度 ≈40–65%**：在满足"poutcome=unknown ∧ month=may ∧ deposit=no" 的客户中，有 40–65% 同时具备"已婚、有房贷、未联系"等属性，关联度中等偏上。
- **提升度 ≈4.6**：该组合出现 "未订阅+特定属性" 的概率是随机出现的 4.6 倍，关联非常强。
- **杠杆度 ≈0.045**：该模式的实际共现率比独立假设高出约 4.5%，在 11,162 条样本中相当可观。
- **卡方 χ^2≈2,300（p≈0）**：强烈拒绝独立假设，统计显著。

实验结果的合理性评价

1. **侧重点偏向"负例"** 虽然实验目标是"挖掘频繁模式并评估"，但得到的最强模式都以 `deposit=no` 为后件，反映了算法优先捕获了占优类（不订阅）的模式。
2. **平衡性影响小** 实际 IR≈1.11（"no":5873 vs "yes":5289），数据较平衡，不应严重偏向多数类。但由于"无违约""无贷款""未联系" 是最常见的几项，算法自然而然将它们排在前面。
3. **覆盖率与可用性** 每条规则覆盖率仅 ~6%，这批规则更适合**刻画一个小众但高度同质**的子群体，而非用于全量客户策略。

应用建议

- **补充“订阅”方向规则** 重新筛选后件中含 `deposit=yes` 的规则，找出支持度、提升度均优秀的模式，才能为营销提供**积极**的客户画像。
- **策略设计** 对于“未订阅”模式：
  - 例如 `{poutcome=unknown, month=may} → deposit=no` 之类，可以帮助你了解哪些客户组合更容易落空，从而避免资源浪费。 对于“已订阅”模式：
  - 将营销重点放在那些 `{poutcome=success}`、`{contact=cellular, month=jun}` 等模式上，实现**精准覆盖**。
- **A/B 测试** 针对“未订阅高风险群体”，尝试替换话术或变更呼叫时间，看是否能打破该模式。
- **定期复盘** 按月重跑挖掘，监测模式稳定性，及时调整最优客户群体的定义。