模型设计说明文档

李施晨

## 1. 问题概述与本项目的概述

### 1.1 问题背景

随着投资产品种类的增多和投资者需求的多样化，为客户提供精准、个性化的基金推荐变得至关重要。有效的推荐系统不仅可以增强客户体验，还能提升资产管理公司的服务质量和市场竞争力。而在不同业务场景下, 新用户的冷启动, 大量且迅速的粗粒度推荐, 和高定制化的精细推荐, 都对我们的模型选型和数据处理提出了不同的要求.

### 1.2 项目目标

本项目旨在开发一个基金推荐系统，该系统能够根据客户的喜好数据、个人特征和基金的特征等信息，为客户推荐最适合他们的基金产品。通过利用先进的机器学习技术，比如基于内容的推荐、因子分解机（FM）模型和深度学习模型等，我们希望能够有效地捕捉用户特征和基金特征之间的复杂关系，从而提供准确的个性化基金推荐。

### 1.3 项目范围

项目的主要任务包括：

**数据探查与预处理**：分析并准备用于模型训练的数据集，包括数据探索、数据清洗、特征工程等。

**模型开发与训练**：选择适合的机器学习模型，本项目中根据不同的业务场景, 选择合适的模型并对模型进行训练。

**模型评估与优化**：利用离线指标评估模型的推荐效果, 并根据评估结果进行必要的调整和优化, 并设计可能的线上评估。

## 2. 数据探索与预处理

### 2.1 划分训练集和测试集

首先, 在本项目的业务背景下, 我们是对一个给定40个特征的客户推荐我们手上的基金, 所以基金数据是背景知识, 而不用计入训练/测试集的划分, 而我们需要对客户数据进行划分.

为了满足模型的训练与评估, 并且避免预测集数据对于模型训练的干扰, 我们在最早的时候对数据集进行分割, 使得训练集 : 测试集 = 4 : 1.

### 2.2 检查冗余特征, 并对日期型特征进行数值化处理

我们首先要理解数据本身的意义, 删去多余的具有完全一致意义的特征, 比如客户数据中的客户公司代码与客户公司名称一一对应, 所以可以删去其中一个. 而基金成立日期, 如果简单地转化为one-hot型向量, 则会丢失本身的时间意义, 所以我们可以将其转化为成立至今的天数.

### 2.3 检查空缺值/重复值

对于基金数据和客户数据, 我们都需要进行空缺值的检查与填充, 这里我们首先要检查每一列特征中空缺值/重复值的占比, 如果超过95%, 那么要考虑删除这一特征, 一方面这是删去了无法利用的信息, 另一方面可以减少计算复杂度.

对于客户数据, 我们删去feature6, feature22, feature26, feature31, feature34, 对于基金数据, 我们删去feature13.

### 2.4 检测并处理异常值

首先我们要对异常值进行检测, 以客户和基金的前三个特征为例, 我们可以使用柱状图和箱形图分析原始数据的分布情况. 基金特征的分布情况如下图所示,

图表, 直方图

描述已自动生成

图表, 箱线图

描述已自动生成

客户特征的分布情况如下图所示,

图表, 箱线图

描述已自动生成图表, 箱线图

描述已自动生成

可以看出原始数据, 尤其是客户特征存在较多的异常值, 所以我们考虑对异常值进行削顶处理, 得到更合理的分布

图表, 直方图

描述已自动生成图表, 箱线图

描述已自动生成

### 2.5 对数值型特征进行标准化处理

由于不同的数值型特征具有不同的尺度, 如果不加处理, 就会对模型计算造成较大的误差或是不必要的训练负担, 所以我们考虑使用Z分数标准化处理它们, 使得它们在保留本身信息的同时, 归于0-1的共同尺度中.

### 2.6 对类别型特征做One-Hot编码

对于客户的客户公司名称、基金的基金公司名称等类别型特征, 我们使用one-hot编码的形式把它们转换为0/1数值向量, 以便后续模型的使用.

### 2.7 构建目标值

对于客户数据的最后四位特征, 第一/二/三/四选择基金, 这是具有喜爱程度高低的特征, 但没有时间上的先后顺序, 所以我们使用评分的形式对它们进行转换, 也就是分别转换成4/3/2/1分, 同时对基金代码进行Multi-Hot编码, 建立rating评分数据, 每行都代表一个客户对各基金的评分, 是前四位选择就填入对应分数, 否则为0. 评分数据的前几行几列如下图所示.

图片包含 日历

描述已自动生成

### 2.8 拼接数值型特征和幻化后的类别型特征

我们将基金数据和客户特征分别做前6步处理后, 对数值型特征和幻化后的类别型特征进行拼接, 获得了新的基金数据和客户数据. 其中客户数据抛去“第一/二/三/四选择基金”成为因变量X1, 基金数据成为因变量X2, 第7步所获得的评分数据矩阵则为应变量Y, 我们所需要研究的模型F就是使得 F(X1, X2) = Y.

处理后的基金和客户数据, 前几行几列如下所示

表格

描述已自动生成

表格

描述已自动生成

### 2.9 相关性分析

对于处理后的数据, 我们还需要对它们的相关性进行检查, 如下图所示, 左侧为客户数据, 右侧为基金数据, 所有非对角线元素均为蓝色, 代表相关性系数较小, 可以排除多重共线性的干扰 .

图表

描述已自动生成图表, 直方图

描述已自动生成

## 3. 模型选型与分析

对于不同的业务场景以及计算资源, 我们需要选取不同的模型.

### 3.1 冷启动

对于刚刚注册的新用户, 我们缺乏他们准确的特征信息, 以及对基金的偏好选择信息, 所以此时可以选择使用content-based方法, 为他们推荐与他们相似的用户偏好的信息. 相似系数结合由第二部分所获得的客户向量之间的距离, 并结合已知客户数据中的偏好评分计算. 基于用户相似度的推荐效果如下图所示, 可见与实际的top选择有所差别.

一些文字和图案

中度可信度描述已自动生成

### 3.2 粗粒度推荐

客户打开基金首页时, 我们需要为客户推荐较大量的基金, 这时可以采用比冷启动更准确的推荐模型, 比如因子分解机, 或者较轻量级的深度学习模型, 一次性给客户快速推荐50款潜在基金. 这一方面达到了一定精度, 另一方面也可以节省计算资源, 加快推荐速度

图表

描述已自动生成

因子分解机FM模型示意图

### 3.3 个性化定制推荐

对于客户打开特定类型基金等需要精细化推荐的场景, 我们可以调用更多的计算资源, 且由于精细化场景中基金数量更少, 我们获得了更多的时间, 所以我们需要考虑使用更精准, 但更耗费计算资源、时间成本的模型, 比如深度学习模型, 这中间较为轻量级的有DeepFM, 而更大型、更准确的模型比如DIEN等, 我们可以在拥有更多客户购买/浏览基金历史信息的情况下使用.

图示

描述已自动生成

DeepFM示意图

## 4. 模型评估

### 4.1 离线评估

除了从数学角度选取了基于连续值的RMSE、MAE, 我们还选取了具有实际意义的基于离散值指标:

准确率（Accuracy）：衡量模型正确推荐用户感兴趣的基金的比例。这样可以通过减少错误的推荐，确保用户看到的基金是他们可能真正感兴趣的, 提高用户体验

召回率（Recall）：衡量模型成功推荐用户感兴趣的基金的比例，考虑了模型漏掉的推荐。高召回率表示模型能够捕捉到更多用户的兴趣，确保不错过可能的好机会

### 4.2 线上评估

如果有机会将模型上线, 那么可以选取:

点击率（CTR）：衡量用户点击推荐内容的比例。CTR衡量了客户对推荐内容的兴趣程度

转化率（Conversion Rate）：我们的目标是促使客户购买基金，所以转化率是一个关键指标，它衡量了成功推荐的比例

## 5 附录

具体数据和可执行代码请查看其他文件.