### 1自我介绍

面试官们好，我叫吕九峦，来自软件工程学院，现在是研究生二年级。字节跳动公司一直是我们非常尊敬的公司，这一次我应聘的岗位是项目管理实习。在研究生阶段，我参与过一个科研项目从了解行业背景开始、到研究题目确立，再到最终论文发表结题的过程，在这个过程中我积累一些项目管理的经验，懂得项目管理的重要性。我对mysql数据的一些基本使用与底层原理也比较熟悉，包括索引底层的结构、事务特性和MVCC机制等。最后，希望能够加入贵公司从事项目管理相关工作，谢谢。

很高兴能够参加中移物联卓越工程师班的面试，我非常欣赏中移物联提出的“为理想，同奋斗”的宗旨，我现阶段的理想就是希望在研究生三年科研和学习的过程中，锻炼我的能力和技术，在三年之后能够进入像中移物联这样优秀的企业。为了这个理想，我也作了很多奋斗，比如，在研一阶段，我就积极参与科研项目，认真解决科研过程中遇到的问题，最终产出了一篇论文，这篇论文投递给科院认定的SCI二区的一个期刊上，编辑给了我们小修意见，最近也是在紧锣密鼓地进行修改。同时，我对专业技术学习也有很高的热情，因此，参加了许多转学学科竞赛，获得了软件测试大赛省赛一等奖和蓝桥杯算法比赛省赛二等奖的奖项。最后，我希望能够成为中移物联的一员

### 2 教育经历

multi-party federated recommendation based on semi-supervised learning 基于半监督学习的多方联邦推荐

2.1 multi-party

两个及以上的参与方

2.2 federated

Federated是联邦的意思，这个词不是机器学习领域专有名词，它来自于政治领域，意思是多个自主政府组成的一个单一主权国家，比如美国，他的政府成为Federated Government 联邦政府。我们熟悉的FBI Federal Bureau of Investigation (FBI)：联邦调查局。在机器学习领域，federated来自federated learning 联邦学习这个改念。联邦学习是一种分布式机器学习方法，它允许模型在用户设备上学习并更新，而无需将用户数据发送到中心服务器，从而保护了用户的数据隐私。

分布式机器学习是一种利用多台计算机（分布在不同的地理位置或在同一网络中）并行处理数据和算法以加快学习速度和处理大规模数据的机器学习方法。

联邦学习分为横向联邦和纵向联邦，这与参与方之间的数据分布有关。举个横向联邦例子，不同地区的两家银行进行合作建模。第一，由于是不同地区，所以两家银行面对的客户也是不同的，如果我们给每个用户一个ID，称用户为一个样本，那么这两家银行的样本ID重叠较少。第二，由于都是银行。那么业务范围都差不多，包括存款、取款、转账、汇款等，数据库表字段相似，我们把一条记录的所有字段称为数据特征，那么这两家银行在数据特征上重叠较多。综上所述：如果联邦参与方样本ID重叠较少，样本特征重叠较多，我们把这样的联邦过程成为纵向联邦

2.3 semi-supervised learning

半监督学习是一种机器学习方法，使用大量的未标记数据和少量的标记数据来训练预测模型

我们都知道，一个典型的机器学习方法模式是把数据和数据对应的标签投喂给模型，让模型来学习数据特征和数据标签之间的关系。这种模式需要大量的高质量标记数据，但在显示情况，标记数据通常是昂贵且稀少的，举个例子，比如说我要做一个算法，识别重庆邮电大学校园内所有的路灯是否损坏，我数据可能是图像或者视频，这是数据可能通过监控就获取，但是我需要人工的给这图像或视频进行标注，判断图像或视频中的路灯是否损坏，如果损坏我就标注为1，没有损坏我就标注为0。这是一个繁琐的过程，所以标记数据是昂贵且稀少的。

半监督学习方法：

基于生成式的模型

我首先来解释一些什么是生成式的模型， 它的英文单词是Generative Model，近几年，由于chatgpt的出现，生成式人工智能这个概念非常的热门，我前段时间甚至在央视新闻上看大关于规范生成式人工智能的文件发布。对我来说，生成式模型或者生成式人工智能就像一个艺术家，你只需要给它提供大量的数据，比如猫的图片，他就能够创造出新的，以前从未见过的作品，比如画出新的猫的图片。生成式的方法通过学习数据本身的信息，内在的分布，基于这个分布来产生新的数据

基于低密度假设，世界非黑即白，没有中间的灰色地带。

半监督学习的假设是有部分少量的标注数据的，那么我就要这些标注数据来学习一个模型，这个模型按道理将效果肯定不好，应为标记数据是少量的。然后将所有得到未标记数据投喂给这个模型，然后模型输出这些未标记数据成为某个类别的概率，由于我做了低密度假设，这个世界非黑即白，那么数据如果属于某一类了，那么属于其他类的概率会非常低，我就可以之间给未标记样本赋予标签。

推荐

推荐系统或者推荐方法，这是一个应用领域。通过用户过去的行为和喜好，预测可能喜欢的商品或服务。我们是从对灵活就业人员推荐公积金来衍生出这篇论文的

### 项目经历（重中之重）

* 1. 灵活就业人员住房公积金制度相关研究

灵活就业人员：就是就业方式相对灵活的人员。与在职职工作不同，他们可能没有固定的工作地点和工作时间。

公积金：全称为住房公积金，相当于一种强制储蓄，个人和公司各交一半，但是所有钱都算在个人账户上，公积金经常用于买房、看病、自己花等用途。

在以前，只有在职员工公可以缴存公积金，重庆是在2021年8月左右开始执行灵活就业人员参加住房公积金的制度

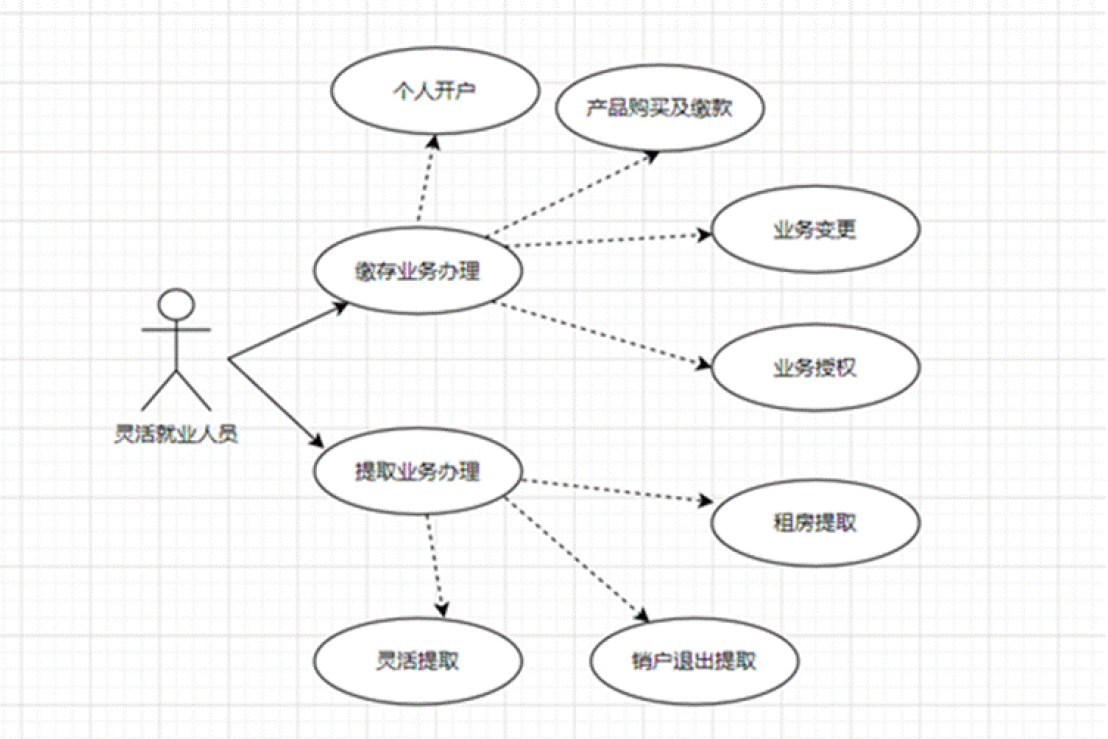
* 1. 通过Navicat和Power Designer等工具，对公积金数据库中与灵活就业人员相关的表进行字段和表分析

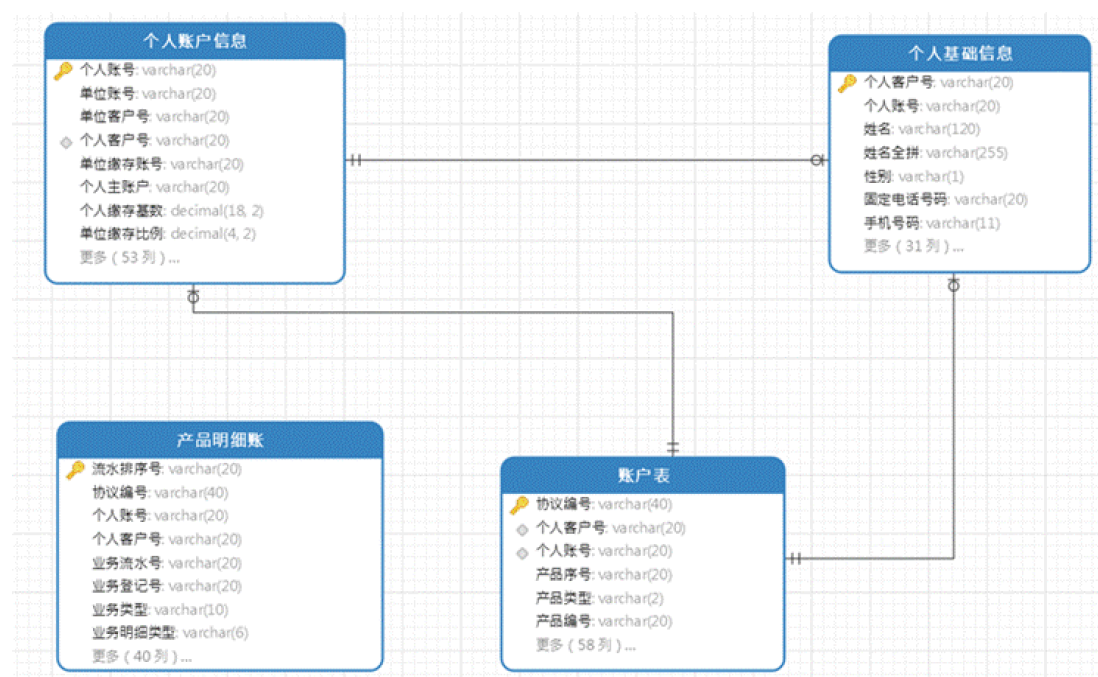
Navicat：

Navicat是一个数据管理工具，支持多种数据库系统，比如mysql，oracle。我主要是用Navicat查看灵活就业人员相关表有哪些字段，这些字段有什么含义，什么数据类型，Navicat也可以可视化不同表之间的关系，一对一、一对多还是多对多，通过对字段和表之间关系的分析帮助后面做特征工程时将字段值编码成数字，删除一些不太必要的特征等等。

Power Designer：

Power Designer是一款建模软件，我在研一上软件建模这门课时学习到的，并把它应用到这个科研项目过程中，我使用Power Designer主要是为了将公积金那边业务文字描述抽象为图，主要是用例图这样有助于我了解行业背景，分析这个业务流程。





用例图：

描述一组用例、参与者及它们之间的关系

参与者：与这个系统相关的人员（用户、组织、外部系统）

用例：对系统功能的描述、通常是动宾短语，用一个椭圆表示。比如灵活就业人员公积金缴存系统的一个用例，参与者是灵活就业人员，那么可以关联两个大的用例，缴纳公积金和提取公积金。

关系包括 关联、泛化（参与者与用例）；包含include、扩展exclude（用例之间）

包含关系：将两个及以上用例共同的行为抽象出来作为以一个用例，比如公积金缴存中授权业务和变更业务都需要登录账号，那么登录账号这个行为就被抽象出来作为单独的一个用例，而授权业务和变更业务与登录账号用例是包含关系

扩展关系：将一个用例拓展，比如查询用户信息，在查询之后发现用户信息错误了，可以扩展以下修改用户信息

泛化关系：也就是继承的这个概念，is-a的关系 比如说普通用户和管理员用户管理员是普通用户的一种特殊形式，可以用泛化关系表示

类图:

类图是一种用于展示系统中类以及它们之间关系的静态结构图。在面向对象的程序设计中，类图是非常重要的工具。在类图中，主要包含以下部分：

类：类是一些具有相同属性（变量）和方法（函数）的对象的集合。在类图中，类通常用一个矩形框表示，框中的上部分写类的名称，中间部分写类的属性，下部分写类的方法。

关系：类与类之间的关系，包括关联（两个类之间有某种联系）、聚合（一个类是另一个类的一部分）、组合（一个类的生命周期依赖于另一个类）和继承（一个类继承另一个类的属性和方法）。

发现未标记样本缺失的PU问题(UDD-PU problem) ，提出联邦学习+PU learning的解决方法

PU问题

PU 是Positive and Unlabelled 缩写，数据的形式只有正样本和未标记样本，从这样的数据形式学习机器学习模型的问题。PU问题在现实中还是很常见的，尤其是推荐领域，举个产品推荐例子，一家电商网站工作，想根据用户的购买历史和浏览行为来推荐产品。我们把用户喜欢的产品称为正样本，用户不喜欢的产品称为负样本。在这场景下，正样本是很容易获取的，比如用户购买过、添加到收藏的、添加到购物车，浏览次数多的产品都可以认为是正样本，但是获取负样本很困难，用户不会明确地去告诉这么多产品中他不喜欢那个产品，那么如何根据现有的部分正样本数据和大量未标记数据，这些未标记数据中可能保护用户喜欢的产品也可能保护用户不喜欢的产品，训练一个模型来预测用户是否会购买某种产品。

现有的推荐方法，比如基于内容的推荐主要关注正样本，他会分析用户的购买历史，来推荐与正样本相似的商品，比如我收藏了一把雨伞，那么他就会给我大量的推送雨伞相关的商品。

而PU方法用于推荐系统领域会不仅利用正样本的信息，也会利用未标记样本的信息来训练模型进而去预测。这是PU方法与其他推荐方法不同的地方

UDD-PU

UDD-PU是PU问题的一种特殊情况，是公积金方法想做灵活就业人员公积金缴存推荐遇到的更为极端的情况，我们把有意愿或者已经购买公积金的人员标记为正样本，对缴纳公积金完全没有意愿的人员标记为负样本。由于缴纳公积金需要用户在支付宝或微信小程序上搜索重庆公积金，选择灵活就业人员缴纳公积金，然后注册个人账户，这时候就会在公积金数据库那边生成一条记录，也就是说公积金数据库表中记录的人员应该至少也是了解过灵活就业人员公积金缴纳，而且有意向缴纳公积金的人员，他才会注册账号，根据我们前面的划分，公积金那边的数据库只包含正样本，传统的PU方法无法解决，因为PU问题需要同时有P Positive和U Unlabelled。这个新的问题我们把它称作UDD-PU问题。

为了解决正负样本数据下的推荐---我们采用PU问题

为了解决UDD，未标记样本的缺失----我们选择了联邦学习。我们要想补充公积金方方的未标记样本，我们想到的解决办法是能不能将社保局、工商局的数据结合起来，与公积金一起合作建模，公积金提供正样本数据，其他参与方提供未标记样本数据，在一个能够保护隐私安全的框架下进行联合建模，隐藏我们选择了联邦学习。

综上所述，这就是联邦学习+PU learning的解决方案提出的过程。

使用draw.io和Visio等作图软件，将研究思路和方法抽象成框架图和算法流程图

PU方法：

直接法，这个方法非常直觉，我们不是有正样本和未标记样本吗，那么我就直接把未标记样本看作样本，训练一个二分类模型。当然，这样训练出来的模型效果不好，因为样本不均衡，正样本数量少，负样本数量太多。

Bagging方法集成

Baggging是集成学习的一个概念，这个方法利用多次随机抽样重构数据集，降低模型的方差，既然直接法会导致样本不均衡的问题，那么我每次从未标记样本中随机采样，采出的样本大小未正样本数据集大小的未标记样本，作为负样本集。把正样本与负样本集合作为训练集，这样就构成了样本平衡的训练集。

2Step方法

2step是另一种解决直接发样本不均衡问题的方法，这个方法主要分为两步，第一步：直接利用直接法训练一个分类器

第二步：将所有的未标记数据投喂给这个分类器，分类器会输出每一个未标记样本未正样本的概率，基于这个概率，我们可以挑选概率最高部分未标记样本，将它们加入到正样本数据集中。然后重复第一步

通过不断扩充正样本数据集，两步法解决了直接法中样本不均衡的问题。

提出VFPU算法，利用Python编程语言，结合scikit-learn、numpy、pandas等机器学习库，实现该算法

你能简单介绍一下VFPU算法的工作原理和主要特点吗？

VFPU让多方以保护隐私的方式协作训练机器学习模型。该算法包括随机抽样、平衡正负样本训练和迭代选择可靠的正样本

你在实现这个算法的过程中遇到了哪些挑战？你是如何克服这些挑战的？

如何将现有的纵向联邦分类器融合到VFPU算法，对相关文献和代码仓库进行调研，选择了fedtree这个代码仓库作为基分类器的实现库

你是如何验证这个算法的有效性的？你有进行任何形式的测试或评估吗？

我们选择了三个公开数据集数据集，分别是银行营销数据集、信用卡违约数据集、成年人人口普查数据集，实验结果表明我们的VFPU算法再解决UDD-PU问题上比其他的联邦半监督方法要好

你能分享一下使用Python，以及scikit-learn、numpy、pandas等库在实现该算法过程中的一些经验和教训吗？

在这个过程总结了很多经验教训，第一，在编码前的准备工作很重要，尤其是整个系统架构的设计、算法流程图的设计，在编码过程中再发现之前设计有问题只能将代码推导重新来过，这样会耗费大量精力，所以前期的设计非常重要。第二，要尽量站在巨人的肩膀上，有时候某块功能实现起来很复杂，自己一个人想了很久的搞不明白，这时候可以进行调研，搜索相应的关键词，看看别人是否已经解决了整个问题，自己能不能利用，总之，不要重复造轮子，对于现有的代码实现重复利用起来，提高效率高。第三就是再编写我一个小模块后测试很重要，要尽可能想出完备的测试用例

这个算法的应用场景是什么？在实际应用中，它的性能如何？

用在只有正样本，未标记样本缺失的推荐领域，性能由于现有的其他联邦半监督学习方法

在实现VFPU算法的过程中，你学到了哪些重要的技能或知识？

第一，学到了数据预处理技术，包括数据清洗、特征编码，对类别型数据进行独热编码，对数值型数据进行标准化。第二，学到了很多机器学习分类模型如逻辑回归、决策树、梯度提升书、随机森林、k近距邻以及如何用sckit-learn来实现这算法。第三，学到了python中处理逻辑代码的技巧，比如迭代器、解包操作、lambda 表达式等等。

如果有机会优化或改进这个算法，你有什么想法或计划？

这个算法仍有改进空间，比如在算法在并行上做的不够好，导致运行时间相对较高。还比如在出里未标记数据的细节上需要改进

你能解释一下这个算法在处理具体数据时的步骤和流程吗？

两层循环，外层循环用来选择可靠正样本，并将其加入到正样本集合中，内层循环用于构建平衡的二分类集合、纵向联邦训练、预测。

你是否考虑过算法的计算效率和资源消耗？在这方面有什么优化措施吗？

加大并行的力度

你是如何保证算法的稳定性和可靠性的？

我会在算法某些位置出进行模型各种参数的保存，如果算法崩溃了，还可以修改后从文件读取参数继续学习

提出三个研究问题，设计验证实验证明VFPU算法的有效性

三个研究问题：

1.多方联邦对于算法的影响？，联邦带了的模型性能降低的影响是否能够忍受。理想情况下，把所有参与方的数据集中到一个数据中心训练机器学习模型是最好的，为联邦虽然能在各参与方保护数据隐私的情况下情绪合作建模，但有可能带带了特征信息丢失，这与联邦联邦学习设计有关，我们希望这种由于特征信息丢失带来的模型性能的损失是可以容忍的。

2.VFPU的基分类器中哪个是最优的？VFPU消融实验，更换不同VFPU的组件，寻找最优的。

3.基线对比试验，VFPU与其他联邦半监督学习方法对比？

数据集：三个公开数据集，银行产品营销数据集、信用卡违约数据集、成人人口普查数据

指标：正确率，准确率，召回率

这篇论文是干什么的？

本文解决了在联邦学习背景下利用多方数据为推荐系统提供服务的难题。提出的VFPU算法能使多方能够在保护隐私的同时协作训练机器学习模型，特别是针对在其他方提供未标记数据的情况下从正样本中学习的问题。

### 技能证书

熟悉python编程

熟悉常见的机器学习分类、集成、降维算法

分类：逻辑回归、感知机、k-紧邻、贝叶斯、决策树

集成：boosting 和 bagging 如果弱分类器是决策树就变成GBDT和RF

降维算法：PCA（priciple component analysis）