目录


**一、个人情况类** ..... 3
(一)自我介绍? ..... 3
(二)为什么从 xxx(主要是实习、工作)离职? ..... 4
(三)为什么觉着自己适合做产品经理 ? ..... 5
(四)为什么想做 AI 产品经理? ..... 6
(五)你做 AI 产品经理的优势? ..... 7
**二、个人经历类** ..... 8
(一)介绍一下你的某段实习、项目经历? ..... 8
(二)你在实习/工作的过程中遇到过哪些困难?最大的困难是什么?怎么解决的?在这过程中你学到了什么? ..... 9
(三)你觉得你实习/工作期间做的最好的项目是哪个?为什么?具体介绍一下?(项目内容)10
(四)你参与/负责的产品在市面上有哪些竞品?你们的竞争优势是什么?你更看好哪款产品? ..... 11
**三、产品素质类** ..... 12
(一)AI 产品经理和传统产品经理有什么区别? ..... 12
(二)AI 产品经理的工作职责和能力要求是什么? ..... 13
1、产品经理对产品方向进行定义 ..... 13
2、产品经理给出产品的设计方案 ..... 13
3、产品经理跟进产品上线 ..... 14
4、产品评估 ..... 14
(三)AI 目前在 B 端和 C 端有哪些落地场景? ..... 14
(四)什么样的 AI 产品算是成功的产品? ..... 15
**四、经典算法类** ..... 16
(一)机器学习和深度学习的关系 ..... 16
(二)机器学习和深度学习的区别 ..... 18
(三)什么是 K 近邻算法 ..... 19
(四)什么是 KNN 算法的实现原理 ..... 20
(五)KNN 算法的优缺点 ..... 20
(六)什么是线性回归算法 ..... 21
(七)线性回归算法的实现原理 ..... 21
(八)线性回归算法的应用场景(广告投放) ..... 22
(九)线性回归算法的优缺点 ..... 23
(十)什么是逻辑回归算法 ..... 24
(十一)逻辑回归算法的优缺点 ..... 24
(十二)朴素贝叶斯算法的实现原理 ..... 25
(十三)朴素贝叶斯算法的优缺点 ..... 26
(十四)朴素贝叶斯算法的应用案例(要不要购买延误险)? ..... 27
(十五)决策树算法的实现原理? ..... 28
(十六)决策树算法的应用案例(预测用户违约)? ..... 29
(十七)决策树算法的优缺点? ..... 29
(十八)什么是决策森林算法? ..... 30
(十九)SVM 算法的实现原理? ..... 31

(二十)SVM 算法的优缺点？ ..... 32
(二十一)K-means 算法实现原理？ ..... 33
(二十二)应用案例：K-means 算法对用户分层？ ..... 34
(二十三)K-means 算法的优缺点？ ..... 36
**五、 深度学习类** ..... 37
(一)什么是神经网络？ ..... 37
(二)什么是 CNN 算法？ ..... 37
(三)CNN 模型的应用场景？ ..... 38
(四)CNN 模型的优缺点？ ..... 38
(五)什么是 RNN 模型？ ..... 38
(六)RNN 模型实现原理？ ..... 39
(七)RNN 模型应用场景？ ..... 40
(八)RNN 模型的优缺点？ ..... 41
(九)什么是 GAN 模型？ ..... 42
(十)GAN 模型实现原理？ ..... 42
(十一)GAN 模型应用场景？ ..... 43
**六、 大数据模型类** ..... 43
(一)什么是大模型？ ..... 43
(二)什么是 ROC 曲线？ ..... 44
(三)什么是 AUC？ ..... 45
(四)什么是 Transformer 模型？ ..... 46
(五)什么是 ChatGPT 模型？ ..... 48
(六)什么是 Diffusion 模型？ ..... 49
**七、 技术基础类** ..... 49
(一)什么是特征清洗、数据交换？ ..... 49
(二)什么是过拟合和欠拟合？ ..... 50
(三)什么是跨时间测试和回溯测试？ ..... 52
(四)什么是训练集、验证集和测试集？ ..... 52
(五)你之前负责产品中使用的最核心的算法是什么？这种算法有哪些优缺点？ ..... 53
(六)你对深度度学习有哪些了解？深度学习的应用场景有哪些？ ..... 53
(七)什么是机器学习？ ..... 54
(八)机器学习的应用场景都有哪些？ ..... 54
(八)逻辑回归相比于线性回归，有什么区别？ ..... 55
(九)你能介绍一下 KNN/朴素贝叶斯/SVM/CNN/Diffusion/NLP 的原理吗？你熟悉哪几种深度学习和机器学习算法？都有哪些区别 ..... 56
**八、工作场景类** ..... 56
(一)AI 算法工程师说你的需求实现不了怎么办？ ..... 56
(二)工作中做的最失败的事情/项目/遇到的最大困难是什么？ ..... 56
(三)请说说你们产品的主要竞品是谁？ ..... 57
(四)如果公司研发资源不足以实现你想要的功能？怎么办？ ..... 57
(五)训练模型时数据集都有哪些来源？ ..... 58
(六)工作中用什么样的方法清洗数据？ ..... 59
(七)模型构建流程通常包括几个阶段？ ..... 59
**九、行业认知类** ..... 60

(一)你怎么看待 AI 或者人工智能行业？对于整个 AI 行业有哪些认知？ .....60
(二)结合我们公司的业务场景？通过 AI 技术可以做哪些工作来提升用户体验？ ..... 60

# 一、个人情况类

## (一)自我介绍?

参考答案:

个人信息+工作经历（实习或在校经历）+岗位关联度（岗位倾向）+结语

个人信息里包含了，姓名，学历 (如果学校比较 nice，可以提一下学校名称)，如果有 HR 在场，最好能带上在本地工作时间，和未来将在本地工作时间，来打消会辞职回老家的顾虑。

工作经历（实习或在校经历），最好是离你比较近的这段工作经历里你觉得最好的那一面拿出来说说，这个时候也可以放个钩子，这个钩子的作用就是自我介绍后，诱导面试官跟着你的节奏走下去，展示你与应聘岗位的匹配度。

比如说，我在 XX 公司负责 XX 产品期间，做了 XXX，用户留存增长 XX%，使用时长提升 XX%，在此经历后，我研究出了自己的一套方法论，在 XXX 方面能够更快找到问题的解决方法。

这个方法论和增长策略就是你的钩子，可以诱导面试官接下来询问更多细节，那就能让面试官跟着你的节奏来，让整个面试更顺畅。

岗位关联度，首先问清楚应聘的岗位和岗位职责。你可以谈谈你对应聘岗位的认识了解，说明你选择该岗位和公司的强烈愿望，可以谈如果被录取，你将怎样尽职尽责地工作，且展示一下在工作经历中，有哪些工作或者经验是和应聘岗位关联度较高。

结语，我的自我介绍完毕，希望与您有一个愉快的面试过程，请面试官提问。

自我介绍的要点：

1.时间要把控好，基本上就是 3~5 分钟必须完结掉，一个好的谈话，应该是说少而听多。

2.可以在自我介绍的时候埋下钩子，比如你的方法论，你觉得好的产品或者功能等。但是千万别给自己埋坑，说的每一个字都需要经得起推敲，千万别被自己带到沟里。还没弄明白什么叫需求分析，就夸夸而谈，很可能被面试官一个问题问懵，那基本就告别了。也包括，离职原因等等，不必要就不要自己把原因什么的就说了。

3.自我介绍要有充分的信心，最好是在面试前，自己先练习几遍。非必要，在自我介绍时，不要谈自己的优点，因为有优点必有缺点，如果缺点环节没准备妥当，可能也是个坑。

## (二)为什么从 xxx(主要是实习、工作)离职?

参考答案:

先了解一下为啥要问这个，第一，是想了解你，通过你的离职原因可以了解你在哪个地方会比较反感，也是了解你下次离职是什么原因。第二，通过了解你的反感区域，去匹配本公司的点，如果本公司真的有这个点，那可能会拒掉（其实也是保护你避坑）。

**回答要点：**

1.要正面回答，不要支支吾吾，更不要答非所问。

刻意回避该问题，会让面试官觉得一定是比较难以启齿的问题，就会联想那些不好的，比如薪资太低想跳槽，不喜欢加班等等内容，往往会做出了不利于面试公司录用你的错误方法。

2.客观说法，不要抱怨。

不要用抱怨的心态数落前公司的不足之处，多从自身出发，陈述原因。比如前公司的薪资低，想跳槽加薪，你可以说，对于自己的职业规划来说，想去一个更大的平台发展和学习等等；公司总加班，你可以说由于身体或者家庭原因不能接受频繁加班，与前公司的业务安排有冲突。

3.最好是能“承”前“启”后。

在阐述离职原因时，可以降低甚至淡化前公司的消极因素，在表述前公司的时候，顺便夸一下当前公司的彩虹屁。比如“在前公司得到了什么赞誉，哪些业绩，但是公司上升空间有限，个人发展受限，正好看到贵公司在招聘，想来贵公司一展身手”。

## (三)为什么觉着自己适合做产品经理？

参考答案:

先明确一下，这个问题适用于 3 年内的产品经理面试，就是初级以内的。超过 3 年，你可以直接回，我在这行 3 年了，业绩也不错，应该适合吧，然后把，上家公司的业绩再说一下。

**这道题是为了:**

1.解应聘者，对产品经理岗位的理解（助理、实习生之类）；

2.了解应聘者对自己的认识，应聘者的能力优势；

3.应聘者的思维逻辑是否条理清楚。

**解题要点，先明确“适合”，后表现“兴趣”。**

1.明确“适合”，就得先有对产品岗位的认知，只有知道这个岗位的职责能力，才会找到自身能力与其适合点。如果连认识都没有，就会牛头不对马嘴，基

本瞬间就 PASS 掉。“适合”可以从这个模板来回答，产品岗位职责能力是 XXXX，自身能力是 XXXXXX，举例子 XXXXX。

【回答实例】：产品经理需要能准确抓住用户的痛点，设计产品去解决用户问题，实现价值，并在公司里也是个承上启下，前后沟通的桥梁，是重要有价值的角色，我在校学生会做活动时就实时发现了什么问题，并通过什么办法，通过与多少个部门进行沟通协调，解决了该问题，得到老师好评。

2.表现“兴趣”，可以通过平时生活对产品的兴趣和研究精神来回答。

【回答实例】：平时脑子里会有很多的想法，个人也比较注重新体验，经常有一些新点子，也比较喜欢互联网，喜欢了解各种互联网产品，看他们解决了什么问题、定位是什么，为什么设计这样的产品功能。如果能自己做出一款改变一些人的产品，一定会很棒！工作中我也乐于思考，能承受一定压力，直面挑战，不断的提升和突破自己。

## (四)为什么想做 AI 产品经理?

参考答案:

第一点是未来必定是 AI 的时代，ChatGPT 开启了 AI 时代纪元，最近二三个月 AI 领域出现了爆发式的发展，还有海内外巨头公司都纷纷下场，未来一定会涌现出大量的 AI 使用场景需求，犹如当年的移动互联网，雷军说过，站在风口上，猪都能飞起来；

第二点，长期主义的考虑，当下很多人都在讨论如何用 ChatGPT 变现，不断地学习各种 AI 工具，我前段时间也是疯狂沉迷使用各种 AI 产品，但时间一久，心里就纳闷：各种 AI 工具迭代太快了，压根学不过来，即使学来了，对我未来有什么帮助？真的变现赚钱吗？其实很多自媒体鼓吹 AI，无非是为了割韭菜，就算 ChatGPT 生成视频脚本，剪映自动生成视频，然后上传到抖音上，最后就能

涨大量的粉丝？醒醒吧，怎么可能？业余的怎么比得过专业的，而且专业的也不一定能挣到钱。所以我冷静下来，开始思考，我该如何真正利用好 AI，如何把 AI 跟眼前的工作结合起来，如何抓住 AI 时代的红利；

第三点，结合我的个人职业和夙愿，一直希望创造出用户最喜爱的产品，觉得贼有成就感，能在时代洪流里留下自己的一点痕迹，是一件很有意义的事情。

结合这三点考虑，我决定转行 AI 产品经理，随着未来 AI 场景需求大量的涌现，AI 产品经理岗位一定会迎来爆发性的需求，而且现在从事 AI 产品经理的人并不多，当需求侧远远大于供给侧，未来 AI 产品经理的溢价空间很大，俗称高薪人士。想明白后，我开始认认真真地从最基础的理论学起，抛弃浮躁的心态，做时间的长期主义者。

## (五)你做 AI 产品经理的优势?

参考答案:

理解人工智能的基础知识和原理，包括机器学习、深度学习、自然语言处理等方面的概念和算法。需要掌握数据分析和数据挖掘的技能，能够有效地管理和利用大量数据来支持产品决策。需要具备产品设计和开发的经验，能够理解用户需求和体验设计原则，能够将 AI 技术应用到产品中，并在产品上实现良好的用户交互。需要了解相关的法律和道德问题，如隐私保护、公平性、透明度等，以确保产品符合合规标准并得到用户信任。需要具备优秀的沟通、协调和领导能力，能够与不同部门的团队进行合作，并管理和指导 AI 项目的开发和推广。

# 二、个人经历类

## (一)介绍一下你的某段实习、项目经历?

参考答案:

这个问题的出发点，是为了挖掘求职者的从业经历中最有“分量”的经历，印象最深一般意味着与众不同，从项目的复杂度，应聘者的解决方法的角度，成果等方面最能察觉到一个人背后真正的能力。

回答要点一，深入了解应聘岗位的职责，公司招人都是希望来了就能干活，所以尽量选择的项目是和该岗位比较契合的，所展现出来的能力也是该岗位所需要的。

回答要点二，回答要有逻辑性和层次感。建议按“项目背景+自己所处环境+项目的挑战+解决方案+最后成果”的框架来解答。

1.通过项目背景和自己所处的环境，展示自己的该端经历的理解和思考。

2.通过表达项目的挑战和自己的解决方案，展示自己在这个经历中具备发现问题，解决问题的能力。

3.讲具体的解法和策略，只提炼大概，不展开细节，成果最好能量化指标。

【回答实例】

我在上一份工作中，参与过 X 个中大型的项目，其中让我印象比较深的项目是“XXXX 项目”，当时基于整个行业 XXXXX 等背景，公司计划 XXXX。

项目启动前，我就主动申请加入该项目，希望负责这个项目的核心 XXXX 工作，因为这不仅能发挥我之前累积的项目经验相关，而且还能让我参与到 XXX 这个新的领域。

当时面临比较大的挑战是如何 XXXX，对此，我通过调研了 XXXX 得到了 1.XX，2.XXX，3.XXXXX 的用户需求，并对此进行了 XXXX，优化迭代了 XXXX，最终达到了项目预期，上线后用户量 XXXX，日活跃人数 XXXX。

回答要点三，尽量带上作品去，且回答时围绕作品来讲解，更容易让面试官信任。

## (二)你在实习/工作的过程中遇到过哪些困难?最大的困难是什么?怎么解决的?在这过程中你学到了什么?

参考答案:

回答前先明确一下这个问题，不管是什么项目，总归是有一些大大小小的困难，所以不要回答没困难。没困难，只会让面试官觉得你在上份工作时，要么天天摸鱼，要么是项目组的边缘人物，未真正做过实事，基本上就是 PASS 掉。

这个问题主要考察面试者:

1.在项目组中的角色和参与程度;

2.在面对问题的时候是怎么解决的，解决思路是什么？；

3.是否有对过去所做工作有总结和复盘。

在产品岗会遇到的问题有但不限于包括，

1.项目的时间压力;

2.人员协调沟通问题（包括团队对产品功能有分歧）；

3.资源不足问题;

4.产品功能上线反馈差;

5.产品优化后效果不明显或者倒退；等等。

你可以自己对号入座，从项目上找出 2~3 个比较困难的点，进行整理。记住，一定要给出复盘结果，最好能有方法论产出。

【回答实例】

在 XXXX 项目中，我有遇到过比较大的困难，当时做 XXX 功能优化时，未遍历到所有场景，导致功能测试时有部分异常问题，经测试人员反馈后我做了 XXX，又做了 XXXX，和团队沟通 XXXX，最后保证了功能的稳定上线，从这次事件后，我整理了产品规划中所有可能会碰到的规范性文档。

## (三)你觉得你实习/工作期间做的最好的项目是哪个?为什么?具体介绍一下?(项目内容)

参考答案:

这个问题主要考察面试者，1、怎么定义最好的项目，如何衡量的。2、面试者的执行力和项目推动力如何。3、面对问题，是怎么思考，怎么拆解的。

建议这样的框架来回答:

背景 (包括项目背景和自己所在项目组角色) -目标-解决方案-项目结果 (数据) 。

选择的项目一定是你参与度比较高的，从头跟到上线有产出的那种。这样你才能将来龙去脉说的清楚。

项目背景和角色分工尽量阐述清楚，在说到解决方案时，这里有个小技巧可以判断自己讲述的是否清楚，那就看面试官听完后是否继续追问你已经讲述的内容，因为这个时候面试官需要对你所描述的项目的复杂度，给出的解决方案有个大概的判断。在陈述项目成果时，要用数据说话，比如 xx 指标，提升了多少，原因是什么等等。

【回答实例】

我认为我做的最好的项目/产品是 xxx，原因有：项目达到并超出预期的数据目标，xxx 指标提升多少；我自己在这个项目中得到了很大提升，沉淀了自己做

产品规划、项目管理、标杆产品打造的方法论，并推广应用到 xxx 项目中和分享给团队成员。

这个项目的背景是 xxx，目标是 xxx，当时我作为产品主负责人亲自去到现场，完整体验 xxx 业务的全流程，拆解流程中遇到的各种问题，并针对问题提出解决方案，将解决方案与行业方案进行对标和优化，与 XXXX 相关方等达成一致，然后规划迭代版本，不断迭代，协同 XXX 相关方对产品进行落地推广，在推广使用过程中不断收集用户反馈和查看产品数据，在大家共同的努力下最终完成 xxx 目标。

## (四)你参与/负责的产品在市面上有哪些竞品?你们的竞争优势是什么?你更看好哪款产品?

参考答案:

考察的是对面试者对产品的熟悉程度和对行业的熟悉程度，也可以从回答的结构来分析，入职后的竞品分析工作时的步骤是否有问题，产出的结果是否满足需求。

步骤一，先选择一个自己比较满意，复杂度较高的产品。

步骤二，阐述产品所在行业背景，行业用户画像，及产品所占市场份额或者产品所处阶段。

步骤三，选择竞品，至少 2 个，因为面试官往往会多问，保守一点 3 个以上。

步骤四，明确分析维度，建议从产品定位、竞品明确宣传的核心优势功能、整体的目标和路径规划、产品的定价策略、产品的销售方式和策略、营销、产品宣传渠道、产品背后的支撑体系、文档及培训支持、研发团队、实力、背景等等进行分析。

步骤五，与你负责的产品做比较，突出产品差异化，和你在差异化上找出的核心竞争力。最好能带上数据结果。比如提升了市场占有率什么的。

# 三、产品素质类

## (一)AI 产品经理和传统产品经理有什么区别?

主要考察候选人除了对通用的产品能力技能之外，是否熟悉 AI 产品的特殊技能和要求。

参考答案:

第一：面试官您好，AI 产品经理作为产品经理，核心职责和底层能力与传统产品经理是一致的，仍然是通过一定的产品方案满足用户的需求从而实现业务目标。

但是它与普通的产品经理，主要存在以下两点不同：第一：实现产品目标的技术手段不同。传统产品经理对接的是研发工程师，需要通过研发工程师的代码，来完成产品的功能实现，那他们使用的就是研发技术。

而 AI 产品经理对接的是算法工程师和研发工程师，需要对接算法工程师完成具体的模型，再对接研发工程师进行工程开发联调和上线。最终，我们得到的产品形态可能是一个 API 接口，没有所谓的页面。

基于这种情况，AI 产品经理除了要懂一些基本的研发技术之外，也需要深入学习算法知识，比如工作中常用到哪些算法，以及它们的实现逻辑等等。

第二：AI 产品经理在与技术人员的协作上与传统产品经理有很大不同。传统产品经理和研发协作时候，只需要提供 PRD 文档，对需求进行讲解，有问题及时提供解答就可以了。

但是 AI 产品经理很难产出一个 ROI(投资回报率)指标明确的 PRD 文档，以及我们和算法同学的沟通也不是一次需求宣讲就能完成的，通常我们需要进行多次的沟通确认，并且在沟通中逐渐清晰对于算法目标范围的设定。

## (二)AI 产品经理的工作职责和能力要求是什么？

主要考察候选人对于 AI 产品经理的工作流程和工作职责有一个全局的了解。如果一个求职者真的从 0-1 做过一款 AI 产品，那么这个问题一定不难，所以这个问题也有助于面试官判断求职者是都有简历造假包装简历的嫌疑。

参考答案：

面试官您好，一个 AI 产品上线的流程大致可以分为，需求定义、方案设计、算法预研、模型构建、模型评估、工程开发、测试上线等几个步骤。这其中，产品经理需要主导的节点有定义产品方向、设计产品方案、跟进产品开发和产品验收评估，AI 产品经理的工作职责主要是在这四个步骤得到体现。

1、产品经理对产品方向进行定义

作为 AI 产品经理，首要的职责都是去定义一个 AI 产品。这包括，搞清楚这个行业的方向，这个行业通过 AI 技术可以解决的问题，这个 AI 产品具体的应用场景，需要的成本和它能产生的价值。这就要求 AI 产品经理除了具备互联网产品经理的基础知识之外，还需要了解 AI 技术的边界，以及通过 AI 技术能够解决的问题是什么。

2、产品经理给出产品的设计方案

完成了产品定义之后，产品经理需要给出产品的设计方案。产品的设计方案会根据产品形态不同而不同，比如硬件和软件结合的 AI 产品，会包括外观结构的设计，机器学习平台的产品需要包括大量的交互设计，模型类的产品（推荐系统、用户画像）更多的是对于模型上线的业务指标的要求。

所以，对于 AI 产品经理来说，此阶段的能力要求为，基本的技术知识是必须要了解的。这些包括基本的统计学概率论知识，主流算法的基本原理和应用场景，以及这些算法可以帮助我们达成什么样的产品目标。

3、产品经理跟进产品上线

产品设计完成之后，就到了工程和算法同学分别进行开发的环节了。在这个过程中，作为 AI 产品需要承担一些项目经理的职责，去跟进项目的上线进度，协调项目资源。

因此，这个阶段产品经理至少要知道模型的构建过程是怎么样的。另外，产品经理还需要知道模型构建过程中，每个节点的产出物，以及它的上下游关系。只有这样，产品经理才可以清楚评估项目进度，遇到需要协调资源的时候，也知道产品在这个阶段需要的是什么。

4、产品评估

产品开发完成之后，产品经理还需要验收产品是否满足业务需求。在这个阶段，产品经理的能力要求是，需要知道如何去评估一个模型，评估模型的指标都有哪些，具体评估的过程是怎么样的，以及评估结果在什么范围内是合理的。

## (三)AI 目前在 B 端和 C 端有哪些落地场景?

参考答案:

1.C 端场景：通用的傻瓜式 AIGC（妙鸭等）为 C 端用户以及兴趣爱好者使用，最终可能呈现：社区（小红书等）、社交（朋友圈等）以及自媒体短视频（抖音、快手等）

2.B 端场景：蚂蚁百灵大模型、月之暗面 Kimi 大模型、昆仑万维天工大模型、知乎知海图 AI 大模型、出门问问序列猴子大模型、面壁智能 Luca 大模型等大模型

在近日通过备案，其中包括三个行业大模型，分别是办公行业的金山 WPS 大模型、教育行业的网易有道子曰大模型和好未来 MathGPT 大模型。

## (四)什么样的 AI 产品算是成功的产品？

参考答案：

一、功能性：一个成功的 AI 产品首先要具备强大的功能性。


它应该能够解决用户的实际问题，并提供精准的答案或结果。比如，一个智能语音助手应该能够听懂用户的指令并作出相应的回应；

一个智能推荐系统应该能够根据用户的个人喜好和行为数据提供个性化的推荐内容等。功能性是 AI 产品最基本也是最重要的特点之一。

二、智能性：

AI 产品的核心在于智能，它应该能够模拟人类的思维和决策过程。智能性可以表现为产品的学习能力和适应能力。一个成功的 AI 产品应该能够不断学习用户的偏好，并根据用户的反馈不断优化和改进。通过不断的迭代和升级，AI 产品能够逐渐变得更加智能，为用户提供更加贴心和个性化的服务。

三、用户体验：

一个成功的 AI 产品应该注重用户体验。它应该能够与用户进行自然而流畅的交流，并能够准确理解用户的需求。好的用户体验不仅体现在产品功能的完善，还

包括界面的友好性和操作的简单性。一个优秀的 AI 产品应该能够通过简洁清晰的界面设计，使用户能够轻松上手并得到满意的使用体验。

四、安全性:

随着 AI 技术的广泛应用，数据安全和隐私保护越来越受到重视。一个成功的 AI 产品应该具备良好的安全性能，保证用户的个人信息不被泄露或滥用。同时，AI 产品应该能够辨别和应对各种安全威胁，确保用户的使用过程安全可靠。综上所述，一个成功的 AI 产品需要具备功能性、智能性、用户体验和安全性等特点。通过不断的技术创新和用户需求的满足，AI 产品能够在市场中脱颖而出，为用户带来更好的体验和便利。

# 四、经典算法类

## (一)机器学习和深度学习的关系

参考答案:

深度学习是机器学习的一种，二者是包含关系，比如机器学习还包含强化学习等诸多其他算法。一般要谈机器学习与深度学习的区别，我们最好是将其视为深度学习与传统的机器学习方法的区别。下面引用花书的两张图，来说明这种包含关系。

人工智能 AI，一种是基于知识的，我们人类专家直接将规则写好，然后让 AI 运行这个规则的流程，例如一般单机游戏中的 AI。另一种就是机器学习，我们不写出规则，而是让机器在数据中学习到这些规则。

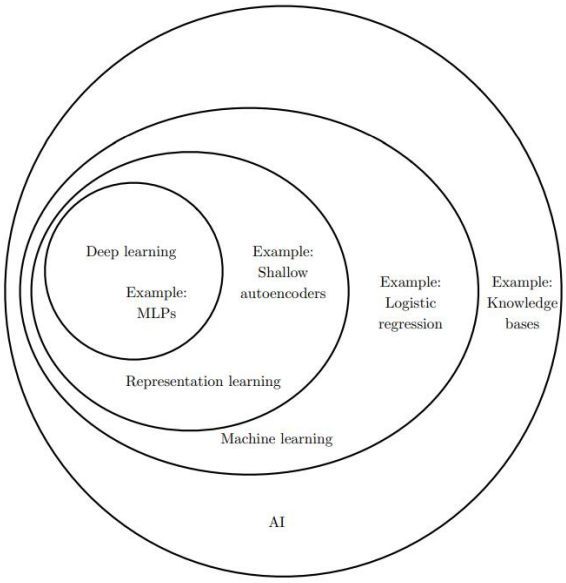



图 1.4: 维恩图展示了深度学习是一种表示学习，也是一种机器学习，可以用于许多（但不是全部）AI 方法。维恩图的每个部分包括一个 AI 技术的示例。

下图中阴影部分是表示每个算法中机器可以学习到的部分，我们从左到右依次进行讲解：

(1) 基于规则的系统，机器没有可以学习的，程序员写好程序交给机器去执行就行了。

(2) 经典的机器学习方法，人类设计好特征，然后将特征送给可学习的特征映射器（分类，回归等）。本篇文章重点讲经典的机器学习方法与深度学习的区别。

(3) 表示学习，对于浅层神经网络，我们只学到简单的特征。对于深层神经网络，开始我们学到简单特征，之后随着网络深度的增加，我们的网络将进一步整合之前的简单特征得到更加高级更加抽象的特征。最右边的这个就是深度学习。

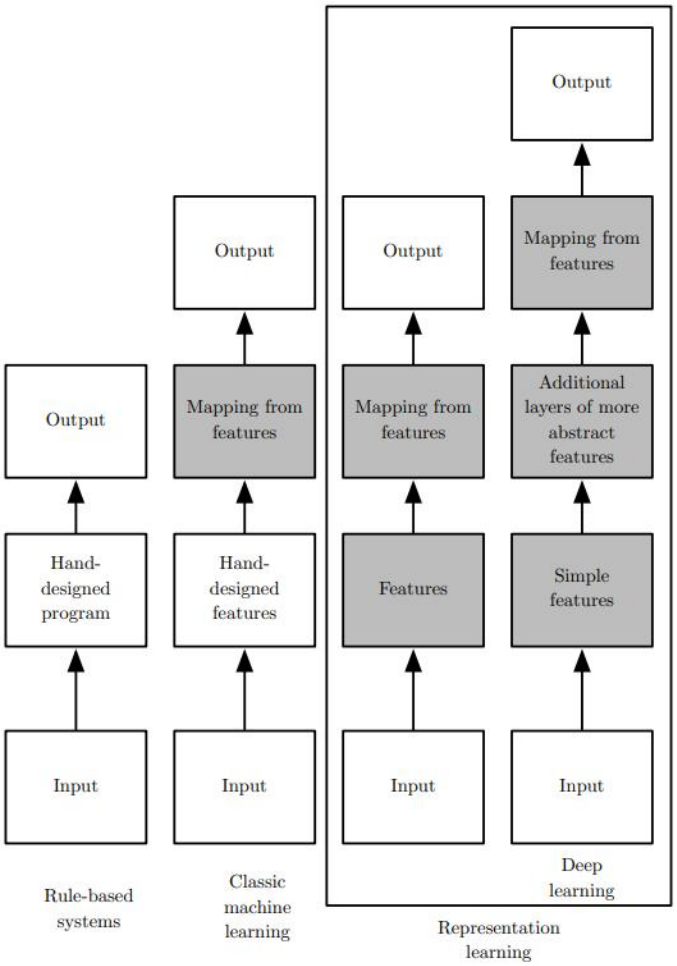



图 1.5: 流程图展示了 AI 系统的不同部分如何在不同的 AI 学科中彼此相关。阴影框表示能从数据中学习的组件。

## (二)机器学习和深度学习的区别

参考答案:

以下将从三个方面来说明二者之间的区别。

(1) 算法流程之间的区别

经典的机器学习方法：输入-->人工抽取特征-->决策树或逻辑回归等算法-->结果。

深度学习：输入-->可学习网络-->结果。

我们可以看到，在深度学习中，并不会人工进行特征的选取，整个过程是一种端到端的学习方式。而在经典机器学习方法中，需要使用人工去精心设计特征，而这些特征第一耗费人力，第二不一定就能找到最适合的哪一个特征。早期用于图像分类的手工特征，例如图像均值，方差，Forstner 算子、SUSAN 算子和 SIFT 算子等。深度学习在算法上有两个优势：精度高，操作简单。

（2）速度上的区别

因为深度学习的模型本身比较大，所以模型需要进行多轮的大量的数据训练，才可以收敛，因此在训练上，深度学习要比经典的机器学习方法要慢很多。但是，在推理阶段，深度学习的执行时间将大大缩短，尤其是在 GPU 的硬件加持下。

（3）数据以及硬件的依赖

深度学习比经典的机器学习方法的模型参数要多得多，因此常常深度学习需要训练的数据也是非常大的，比如 imagenet 这样大的数据集加速了深度学习的发展。另一方面，想要训练的快，就需要使用运算速度非常快的 GPU，因为无论卷积还是 transformer 两者在网络结构上都是可以并行运算的，这正好符合 GPU 的特性。一般使用 GPU 可以加速几十倍的速度，相比于 CPU。数据的增加和硬件的飞速发展也促使了近年来深度学习的关注量要远远超过经典的机器学习方法。

## (三)什么是 K 近邻算法

参考答案：

K-近邻算法(K Nearest Neighbor)又叫 KNN 算法，指如果一个样本在特征空间中的 k 个最相似的样本中的大多数属于某一个类别，则该样本也属于这个类别。也就是对于新输入的实例，从数据集中找到于该实例最邻近的 k 个实例，那么这 k 个实例大多数属于某一个类，那么就把该实例放到该类中。

KNN 算法不仅可以用于分类，还可以用于回归。通过找出一个样本的 k 个最近邻居，将这些邻居的属性的平均值赋给该样本，就可以得到该样本的属性。

## (四)什么是 KNN 算法的实现原理

参考答案:

存在一个样本数据集合，也称作为训练样本集，并且样本集中每个数据都存在标签，即我们知道样本集中每一个数据与所属分类的对应关系。输入没有标签的新数据后，将新的数据的每个特征与样本集中数据对应的特征进行比较，然后算法提取样本最相似数据(最近邻)的分类标签。一般来说，我们只选择样本数据集中前 k 个最相似的数据，这就是 k-近邻算法中 k 的出处，通常 k 是不大于 20 的整数。最后，选择 k 个最相似数据中出现次数最多的分类，作为新数据的分类。

## (五)KNN 算法的优缺点

参考答案:

优点

1.简单好用，容易理解，精度高，理论成熟，既可以用来做分类也可以用来做回归；

2.可用于数值型数据和离散型数据；

3.训练时间复杂度为 O(n)；无数据输入假定；

4.对异常值不敏感

缺点

1.计算复杂性高；空间复杂性高；

2.样本不平衡问题（即有些类别的样本数量很多，而其它样本的数量很少）；

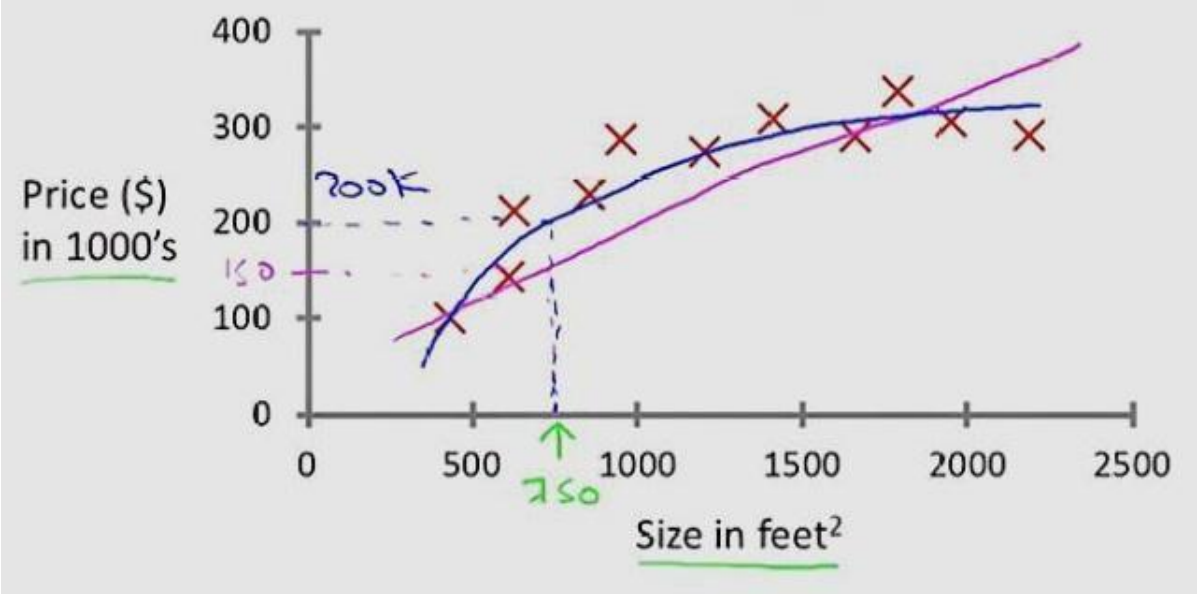
3.一般数值很大的时候不用这个，计算量太大。但是单个样本又不能太少，否则容易发生误分。

4.最大的缺点是无法给出数据的内在含义。

## (六)什么是线性回归算法

参考答案:

线性回归是利用数理统计中的回归分析，来确定两种或两种以上变量间相互依赖的定量关系的一种统计分析方法，运用十分广泛。线性回归模型是相对简单的回归模型，对一个或多个自变量之间的线性关系进行建模，可用最小二乘法求模型函数。




回归分析中，只包括一个自变量和一个因变量，且二者的关系可用一条直线近似表示，这种回归分析称为一元线性回归分析。如果回归分析中包括两个或两个以上的自变量，且因变量和自变量之间是线性关系，则称为多元线性回归分析。

## (七)线性回归算法的实现原理

参考答案:

线性回归算法的目的是来找到一条函数表达式，从而能够最好的拟合给定的数据集。

当得到这条函数表达式的时候，讲所有已知的点代入到这个函数表达式之间，就会得到一个函数值，而这个函数值减去真实值之后就会得到一个误差，将所有的误差平方后求和就是这个函数表达式整体的一个误差，这种将所有误差平方求和的方法叫做残差平方和。得到的结果越大，就说明预测值和实际值差距越大，得到的结果越小，就说明预测值和实际值差距越小，当结果为 0 时，证明所有的点都在这个函数表达式上。

$$y = k_0 + k_1 x_1 + k_2 x_2 + ... + /varepsilon$$

举个例子，在二维平面上一些点随机的分布在一条线的两侧，那么线性回归的目的就是找到这条线的函数表达式，得到这条线的函数表达式就能处理未知的点，就能够求出他的结果值。

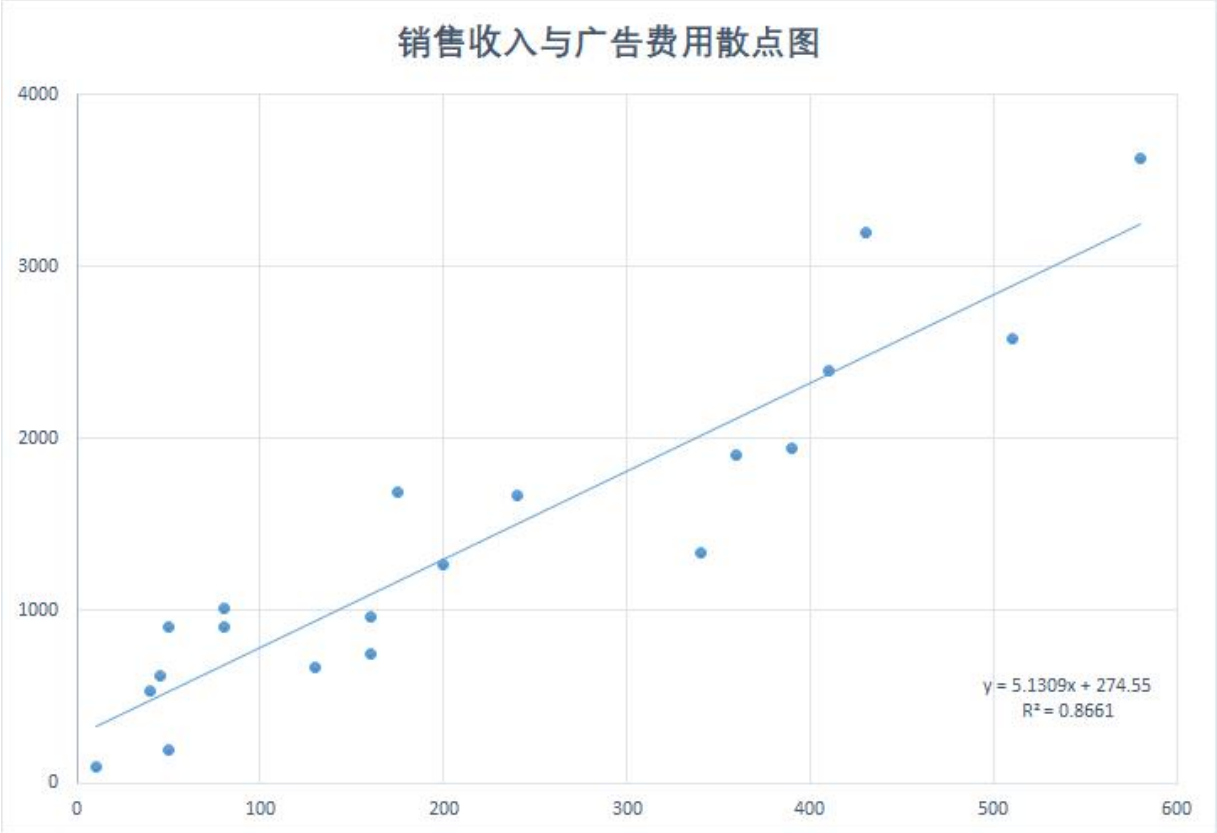
## (八)线性回归算法的应用场景(广告投放)

参考答案：

假如，现在部门要推出一款产品。为了让产品卖得更好，就到处去投放广告，让大家都知道这个产品，激发大家购买的欲望。因为一般来说，广告投放得越多，钱花得越多，知道的人越多，产品卖得越多。

那根据历史累计的广告投放经费和销售额，我们可以画出一张关系图，图上每个点对应的 X 轴代表广告费，Y 轴代表销售额。结合这张图我们可看出，有些坐标点的收益相对较高，有些坐标点的收益相对较低，大概率它们是符合线性关系的。

已知线性回归方程是 Y = AX + B，将已有数据代入到这个方程中，然后求得出一组 A 和 B 的最优解，最终拟合出一条直线，使得图中每个点到直线的距离最短，也就是上面说的损失函数最小。这样，我们就能通过这个最优化的 A 和 B 的值，进行估算广告经费和销售额的关系了。




## (九)线性回归算法的优缺点

参考答案：

优点：

1.运算速度快，由于算法简单，符合非常简洁的数学原理，所以线性回归算法不管时建模速度还是预测速度都是非常快的。

2.可解释性很强，由于最终可以得到一个数学函数表达式，根据计算出的系数就可以明确的知道每个变量的影响大小。

3.善于获取数据集中的线性关系。

缺点：

1.预测的精确度较低，由于获得的模型只是要求最小的损失，而不是数据良好的拟合，所以精确度比较低。

2.不相关的特征会影响结果，对噪声数据处理比较难。

3.不适用于非线性数据。

4.容易出现过拟合，尤其是数据量不大的情况。

## (十)什么是逻辑回归算法

参考答案：

简单来说， 逻辑回归（Logistic Regression）是一种用于解决二分类（0 or 1）问题的机器学习方法，用于估计某种事物的可能性。比如某用户购买某商品的可能性，某病人患有某种疾病的可能性，以及某广告被用户点击的可能性等。 注意，这里用的是“可能性”，而非数学上的“概率”，logisitc 回归的结果并非数学定义中的概率值，不可以直接当做概率值来用。该结果往往用于和其他特征值加权求和，而非直接相乘。

## (十一)逻辑回归算法的优缺点

参考答案：

优点：

(1)对率函数任意阶可导，具有很好的数学性质，许多现有的数值优化算法都可以用来求最优解，训练速度快；

(2)简单易理解，模型的可解释性非常好，从特征的权重可以看到不同的特征对最后结果的影响；

(3)适合二分类问题，不需要缩放输入特征；

(4)内存资源占用小，因为只需要存储各个维度的特征值；

(5)直接对分类可能性进行建模，无需事先假设数据分布，避免了假设分布不准确所带来的问题

(6)以概率的形式输出，而非通过知识直接判断是 0 还是 1，对许多利用概率辅助

决策的任务很有用

缺点：

(1)不能用逻辑回归去解决非线性问题，因为 Logistic 的决策面是线性的；

(2)对多重共线性数据较为敏感；

(3)很难处理数据不平衡的问题；

(4)准确率并不是很高，因为形式非常的简单(非常类似线性模型)，很难去拟合数据的真实分布；

(5)逻辑回归本身无法筛选特征，有时会用 gbdt 来筛选特征，然后再上逻辑回归。

## (十二)朴素贝叶斯算法的实现原理

参考答案：

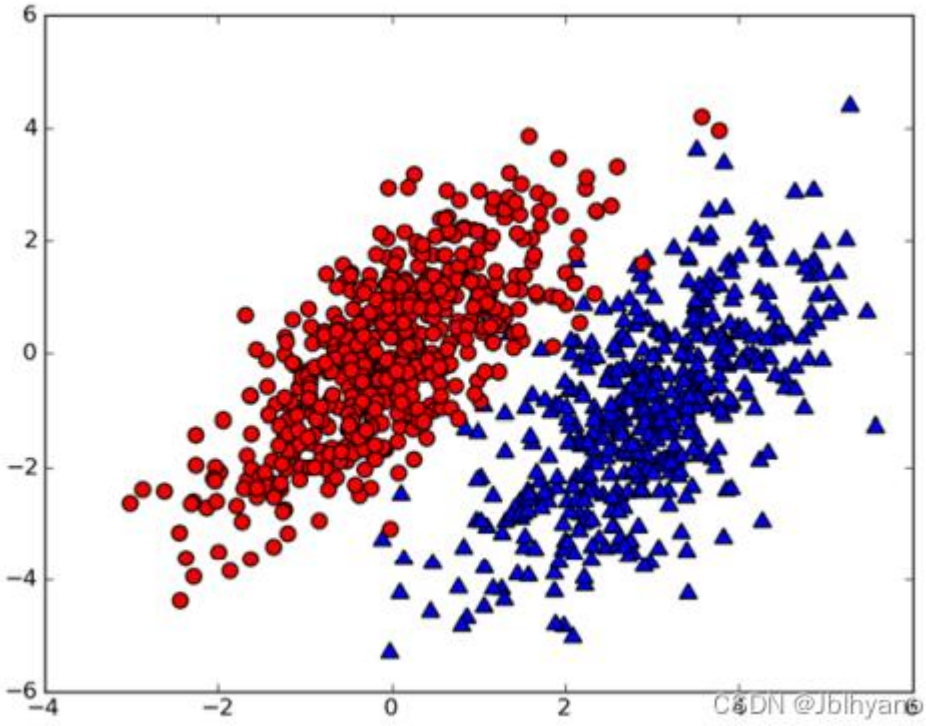
朴素贝叶斯（NB）是一种基于贝叶斯定理和特征条件独立假设的分类方法。本质上朴素贝叶斯模型就是一个概率表，其通过训练数据更新这张表中的概率。为了预测一个新的观察值，朴素贝叶斯算法就是根据样本的特征值在概率表中寻找最大概率的那个类别。

之所以称之为「朴素」，是因为该算法的核心就是特征条件独立性假设（每一个特征之间相互独立），而这一假设在现实世界中基本是不现实的。

简单来说，朴素贝叶斯分类器假设在给定样本类别的条件下，样本的每个特征与其他特征均不相关，对于给定的输入，利用贝叶斯定理，求出后验概率最大的输出。

朴素贝叶斯的基本思想：如果一个事物在一些属性条件发生的情况下，事物属于 A 的概率>属于 B 的概率，则判定事物属于 A。

假设我们有一个数据集，它是由两类数据构成，数据分布如下图所示：




p1(x,y)表示数据点(x,y)属于类别 1（图中用圆点表示的类别）的概率

p2(x,y)表示数据点(x,y)属于类别 2（图中用三角形表示的类别）的概率

那么对于一个新数据点(x,y)，可以用下面的规则来判断它的类别：

如果 p1(x,y) > p2(x,y)，那么类别为 1。

如果 p2(x,y) > p1(x,y)，那么类别为 2。

也就是说，我们会选择高概率对应的类别。这就是贝叶斯决策理论的核心思想，即选择具有 最高概率的决策。

## (十三)朴素贝叶斯算法的优缺点

参考答案：

优点：

1. 算法简单且易于实现：朴素贝叶斯算法基于简单的概率统计原理，模型参数估计简单，算法实现相对容易。

2. 高效性：朴素贝叶斯算法具有高效的训练和预测速度，适用于大规模数据集。

3. 对小样本数据有效：朴素贝叶斯算法在处理小样本数据时表现良好，因为它通过特征条件独立假设来估计参数，减少了参数估计的不确定性。

4. 对缺失数据友好：朴素贝叶斯算法能够处理缺失数据，通过忽略缺失特征的条件概率来进行分类预测。

缺点：

1. 特征条件独立假设过于简单：朴素贝叶斯算法假设特征之间相互独立，这在实际应用中并不总是成立。特征之间的相关性可能会导致分类结果的偏差。

2. 对输入数据分布的假设：朴素贝叶斯算法假设特征的分布满足条件独立性，但在实际情况中，特征之间的关系可能是复杂的，导致模型的偏差。

3. 处理连续特征困难：朴素贝叶斯算法对连续特征的处理相对困难，通常需要进行离散化处理，这可能会导致信息损失。

4. 类别之间的类别比例影响结果：朴素贝叶斯算法对类别之间的类别比例敏感，如果训练样本中某个类别的样本数量远大于其他类别，可能会导致分类结果的偏差。

## (十四)朴素贝叶斯算法的应用案例(要不要购买延误险)?

参考答案：

最近看到一则新闻，王女士从 2015 年开始，凭借自己对航班和天气的分析，成功地购买了大约 900 次飞机延误险并获得延误赔偿，累计获得保险理赔金高达 300 多万元。那么她是怎么决定要买延误险的呢?

其实，航班延误最主要的原因就是天气变化，包括起飞地及降落地的天气；除此之外，也有机场和航空公司的原因。假设这些原因之间并没有互相影响，每一项对于飞机最终是否延误的影响都是独立的，王女士集齐过去的数据，就可以计算出每一个条件与飞机延误的概率。比如，在总体上延误的概率为 20%，不延误

的概率为 80%。在飞机延误的情况下，“起飞地天气=晴天”的概率为 20%，“降落地天气=雨天”的概率为 40%，“机场=首都机场”的概率为 35%，“航空公司=南方航空”的概率为 5%；在不延误的情况下，这些属性的概率分别为 60%、55%、45%、55%。

那么这个时候，有一架南方航空公司的航班，从北京飞往上海，北京天气是晴天，上海天气是雨天，那么，我们就可以根据上面的概率算出来不延误的综合概率=80% × 60% × 55% × 45% × 55% = 0.0065412，延误的综合概率=20% × 20% × 40% × 35% × 5% = 0.00028，从这个结果来看，不延误的可能性要高于延误的可能性，所以这次不需要买延误险。

## (十五)决策树算法的实现原理?

参考答案:

决策树算法的原理是根据已知数据集的特征和决策树算法是.种常用的机器学习算法，它可分类结果，构建一颗树形结构，通过对待分类样才进行特征比较和分类判断，实现对新样本的分类预测

决策树算法的基本原理是根据信息熵和信息增齐对数据集进行划分，构建一棵树形结构。在决策树中，每个节点代表一个特征，每个分支代表这个特征的一个取值，个叶子节点代表一个分类结果

信息简是度量信息不确定性的一种方法，它的作越大，表示信息的不确定性越高。在决策树算法中，我们希望通过划分从而提高分类的准确性。信息增益是指在某个特征上划分数据集前后数据集，让信息熵减少，即让信息不确定性降低信息熵的减少量。我们希望选择信息增益最大的毕征作为当前节点的划分标准，从而构建决策树。

## (十六)决策树算法的应用案例(预测用户违约)?

参考答案:

银行客户流失是指银行的客户终止在该行的所有业务并销号。但在实际运营中，对于具体业务部门，银行客户流失可以定位为特定的业务终止行为。商业银行的客户流失较为严重，流失率可达 20%。而获得新客的成本是维护老客户的 5 倍。因此，从海量客户交易数据中挖掘出对流失有影响的信息，建立高效的客户流失预警体系尤为重要。

客户流失的主要原因有：价格流失、产品流失、服务流失、市场流失、促销流失、技术流失、政治流失。有些时候表面上是价格导致的客户流失，但实际上多重因素共同作用导致了客户的流失。比如说，不现实的利润目标、价格结构的不合理、业务流程过于复杂、组织结构的不合理等等。维护客户关系的基本方法：追踪制度，产品跟进，扩大销售，维护访问，机制维护。

因此建立量化模型，合理预测客群的流失风险是很有必要的。比如：常用的风险因子，客户持有的产品数量、种类，客户的年龄、性别，地理区域的影响，产品类别的影响，交易的时间间隔，促销的手段等等。根据这些因素及客户流失的历史数据对现有客户进行流失预测，针对不同的客群提供不同的维护手段，从而降低客户的流失率。

## (十七)决策树算法的优缺点?

参考答案:

1.决策树所产生的预测规则的形式为：如果 x r 1 ∈ A 1 …且 x r m ∈ A m，那么 Y ＝y，很容易解释。

2.在树的生长过程中，对定序或连续自变量而言只需使用变量取值的大小顺序而不使用具体取值。因为对这些自变量进行任何单调增变换（例如，取对数）都不

改变变量取值的大小顺序，而对自变量进行任何单调减变换（例如，取倒数）把原来取值的大小顺序完全颠倒；所以这些变换都不会改变划分的结果。因此，在建立决策树时，无需考虑自变量的转换（但注意，有时需要考虑因变量的转换）。

3.因为决策树只使用了定序或连续自变量取值的大小顺序，它对自变量的测量误差或异常值是稳健的。

4.决策树能够直接处理自变量的缺失值。如第二章所述，如果数据中有多个自变量存在缺失，决策树可用来插补这些自变量的缺失值。

5.决策树可以用作变量选择的工具。

缺点：

1.每个非叶节点的划分都只考虑单个变量，因此很难发现基于多个变量的组合的规则。例如，可能按照 2 x 1 +3 x 2 的值划分比较好，但决策树只会考虑按照 x 1 或 x 2 的值进行划分，很难发现这样的组合规则。

2.为每个非叶节点选择最优划分时，都仅考虑对当前节点划分的结果，这样只能够达到局部最优，而无法达到全局最优。

3.正因为决策树是局部贪婪的，树的结构很不稳定。例如，若将学习数据集随机分割为不同的训练数据集和修正数据集，可能对于某次分割， x r 1 被选作根节点的划分变量，而对于另一次分割， x r 2 （ r 2 ≠ r 1 ）被选作根节点的划分变量，之后继续划分下去，这两棵树的结构差异会非常大。这种差异也可能使得两棵树的预测性能存在很大差异。而这些差异仅仅是由学习数据集随机分割的差异带来的！此外，因为不同结构的树隐含的预测规则存在不同的解释，所以这种结构不稳定性也降低了决策树的可解释性。

## (十八)什么是决策森林算法？

参考答案：

随机森林是一种基于决策树的集成算法。随机森林通过构建多个决策树来进行分类，每个决策树都是基于随机抽样的训练数据集构建的。在构建每个决策树的过程中，我们会随机选择一个特征子集来进行特征选择，并在每个节点上选择最佳的特征进行分割。通过集成多个决策树的预测结果，随机森林可以减少过拟合的风险，并提高模型的泛化能力

## (十九)SVM 算法的实现原理?

参考答案：

SVM（支持向量机）是一种用于分类和回归分析的机器学习算法。它基于构建一个最优的超平面，可以将不同类别的数据分隔开来，从而实现分类。

具体来说，SVM 的算法原理如下：

寻找最优的超平面：在给定的训练数据中，SVM 算法会寻找一个最优的超平面，使得将数据分为两个类别的间隔最大化。

核函数的应用：对于非线性分类问题，SVM 采用核函数将数据映射到高维空间中，使得在该空间中可以使用线性超平面分割数据。

求解优化问题：SVM 通过求解一个凸二次规划问题来确定最优的超平面。该问题的目标是找到一个最小的误分类率，并最大化分类边界的间隔。

支持向量的确定：在确定最优的超平面后，SVM 算法将寻找支持向量，即离最优超平面最近的训练数据点。这些数据点在分类过程中起到了关键的作用。

分类器的构建：基于最优的超平面和支持向量，SVM 可以构建一个分类器，用于对新的数据进行分类。

总体来说，SVM 是一种强大的分类器，可以处理线性和非线性分类问题。它的核心思想是最大化分类边界的间隔，并利用支持向量来确定最优的超平面，从而实现高效的分类。

## (二十)SVM 算法的优缺点?

参考答案:

优点

1. 能够处理高维数据

SVM 算法的核心思想是将数据映射到高维空间中，使得数据在该空间中更容易被分离。这种映射方式可以通过选择不同的核函数来实现，例如线性核、多项式核、高斯核等。因此，SVM 算法能够处理高维数据，不受维度灾难的影响。

2. 具有较强的泛化能力

SVM 算法采用结构风险最小化原则进行模型选择，即在保证训练误差最小的同时，尽可能地减小泛化误差。这种原则能够有效地避免过拟合现象的发生，使得 SVM 算法具有较强的泛化能力。

3. 适用于小样本数据

由于 SVM 算法采用间隔最大化原则进行分类，因此其分类效果不仅与训练样本的数量有关，还与训练样本的分布情况有关。当训练样本数量较小时，SVM 算法能够更好地处理数据分布不均匀的情况。

4. 可以处理非线性问题

SVM 算法通过核函数的选择，可以将非线性问题转化为线性问题进行处理。例如，通过选择高斯核函数，可以将数据映射到无限维空间中，从而实现对非线性问题的分类。

5. 具有较好的鲁棒性和可解释性

SVM 算法对异常点的鲁棒性较好，可以有效地避免异常点对分类结果的影响。此外，SVM 算法的分类结果具有较好的可解释性，能够清晰地描述不同类别之间的区别。

缺点

1. 对参数的敏感性

SVM 算法中存在多个参数需要进行调节，例如核函数的选择、正则化参数的选择等。这些参数的选择对分类结果有较大的影响，需要进行反复试验和调整。如果参数选择不当，可能会导致分类效果较差。

2. 计算复杂度高

SVM 算法的计算复杂度较高，尤其是对于大规模数据集和高维数据集，计算时间和计算空间都会很大。此外，SVM 算法的训练过程需要多次迭代，也会增加计算的复杂度。

3. 对数据的缩放敏感

SVM 算法对数据的缩放敏感，如果数据没有进行归一化处理，可能会导致分类结果的偏差。

4. 对噪声数据敏感

SVM 算法对噪声数据敏感，如果数据中存在噪声数据，可能会导致分类结果的偏差。因此，在使用 SVM 算法进行分类之前，需要对数据进行预处理，去除噪声数据。

5. 仅适用于二分类问题

SVM 算法仅适用于二分类问题，对于多分类问题需要进行多次二分类处理。此外，对于不平衡的数据集，SVM 算法可能会出现分类偏差的问题。

## (二十一)K-means 算法实现原理?

参考答案:

K-means 是一种常用的聚类方法，它将数据划分为 K 个相似的簇，其中每个簇的中心为该簇内所有数据点的均值。以下是 K-means 的基本原理和步骤:

原理： K-means 基于一个简单的想法：相似的数据点应该在空间中彼此靠近，并且可以通过计算每个点到各个簇中心的距离来找到这些点的簇标签。

步骤：

初始化：首先选择 K 个数据点作为初始的簇中心。这可以是随机选择，也可以是使用某种启发式方法。

分配数据点：对于数据集中的每一个数据点，计算其到 K 个中心的距离，并将其分配到距离最近的中心所在的簇。

更新簇中心：对于每一个簇，计算簇中所有数据点的均值，将均值作为新的簇中心。

收敛判断：比较新的簇中心与上一次迭代的簇中心，如果簇中心没有（或只有微小的）变化，算法结束。否则，返回第 2 步。

结束：当簇中心不再变化或达到预定的迭代次数时，算法结束。

需要注意的是，K-means 的结果可能会受到初始中心的影响，导致局部最优。为了获得更好的聚类结果，通常会多次运行算法，每次使用不同的初始中心，然后选择最好的结果。

## (二十二)应用案例：K-means 算法对用户分层？

参考答案：

RFM 模型是衡量客户价值和客户创利能力的重要工具和手段。根据美国数据库营销研究所 Arthur Hughes 的研究，最近一次消费时间间隔（Recency），消费频率（Frequency），消费金额（Monetary），这三个要素构成了数据分析最好的指标，通过这 3 个指标对用户进行分类，根据不同类别的用户进行精准营销。

最近一次消费时间间隔（Recency）：近度，最近一次有效订购订单距离当前时间点的时间。

理论上最近一次购买的顾客越近越是优质客户，最近才购买商品或服务的顾客，是最有可能再次购买商品或服务的客户，对即时提供的商品或者是服务也最有可能有反应；

最近一次消费的过程是持续变动的，客户的最近一次消费时间间隔会随着时间的变化以及客户的购买行为变化而变化；

最近一次消费时间间隔可以帮助监控业务的健康程度。比如，月报告中显示上一次购买很近的客户(最近一次消费为 1 个月)人数环比增加，则表示该业务是个稳健成长的业务。相反，如上一次购买很近的客户（最近一次消费为 1 个月）人数环比降低，则表示该业务走向衰落的先兆；

消费频率（Frequency）：频度，客户在限定时间内订购订单的次数。

消费频次高的客户，往往也是满意度最高的客户；

根据消费频次，可以把客户分成不同层级，观察用户在不同层级的分布情况，通过运营手段提高消费频次，增加高层级客户占比；

消费金额（Monetary）：值度，客户在限定时间内订购订单的总支付金额。

消费金额是衡量客户价值的支柱指标，”帕雷托法则”——公司 80%的收入来自 20%的顾客，对有价值的客户进行营销能得到更可观的经验效果；

以客户订购订单的 Recency、Frequency、Monetary 来替代客户使用的 Recency、Frequency、Monetary，主要有以下几点原因：

电信行业的客户每天都在使用电信业务的情况下，其最近时间间隔为零，不同的客户区分度很小，客户订购的时间间隔较大，以订购近度替代使用近度，避免了客户使用的近度难于区分的问题。

如果客户在一定时期内使用电信业务的次数数量非常大,则客户的频度也将是一个很大的数量，客户订购的次数相对较少，可以减少统计客户使用次数的工作量。

客户订购支付金额跟客户实际使用消费金额最终是相等的，因此，从订购交费角度构建的 RFM 模型是可取的。

因此需要从客户交费角度来考虑对客户进行 RFM 模型建模，以 RFM 模型为基础,通过客户的 RFM 行为特征衡量分析客户忠诚度与客户内在价值。

从公司所有的客户记录中选择近 2 年内还有消费订购记录的客户进行分析。把这 3 个指标（R、F、M）按价值从低到高排序，并把这 3 个指标作为 XYZ 坐标轴，大于（等于）总 RFM 平均值的为价值高坐标、小于总 RFM 平均值的为价值低坐标。

## (二十三)K-means 算法的优缺点？

参考答案：

优点：

1.实现简单

2.对于大数据集，算法是高效的

缺点：

1.结果可能会受到初始中心选择的影响，导致局部最优

2.对于簇的形状和大小敏感（例如，对于非凸形状的簇，K-means 可能不会很好地工作）

3.需要预先指定 K 值，这在实际应用中可能不容易确定

为了解决部分缺点，有很多变种和改进的方法，例如 K-means++（用于更好的初始化中心）、二分 K-means、DBSCAN（不需要预先指定簇的数量，并且可以发现任意形状的簇）等。

# 五、深度学习类

## (一)什么是神经网络?

参考答案:

神经网络是一种模拟人脑神经系统的计算模型，它由大量的人工神经元组成，通过模拟神经元之间的连接和信息传递来实现复杂的信息处理和学习。神经网络具有自适应性、非线性和并行处理等特点，被广泛应用于机器学习、模式识别、数据挖掘等领域。

神经网络的基本组成单位是人工神经元，也称为节点或神经元。每个神经元都有多个输入和一个输出，输入通过带有权重的连接传递给神经元，神经元对输入进行加权求和，并经过一个激活函数处理后输出结果。神经网络的结构由多个神经元以层次化的方式连接而成，层与层之间的神经元之间存在连接，信息通过这些连接从输入层传递到输出层。

神经网络的学习过程是通过调整连接权重来实现的。在训练过程中，神经网络接收一组已知的输入和对应的输出，通过计算实际输出与期望输出之间的误差，并利用误差反向传播算法来更新连接权重。通过反复迭代训练，神经网络能够逐渐优化权重，提高对输入数据的处理和泛化能力。

## (二)什么是 CNN 算法?

参考答案：

## (三)CNN 模型的应用场景？

参考答案：

CNN（Convolutional Neural Network）是深度学习中的一种前馈神经网络，应用范围广泛，包括图像识别、语音识别、自然语言处理等领域。其主要特点是通过权值共享和池化操作来减少训练参数并提高模型的鲁棒性和泛化能力，可有效地提高识别准确率。

## (四)CNN 模型的优缺点？

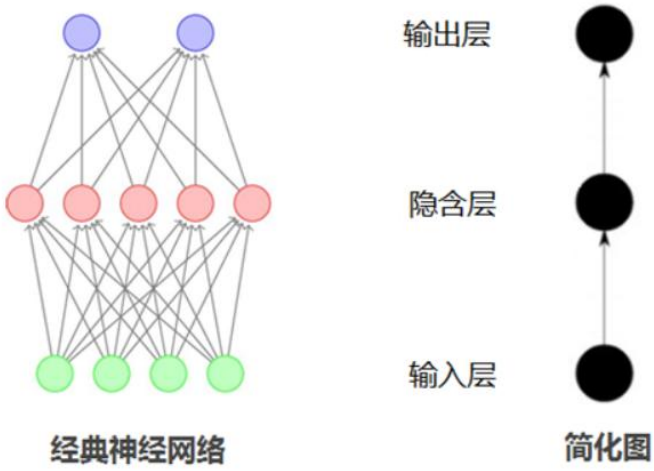
参考答案：

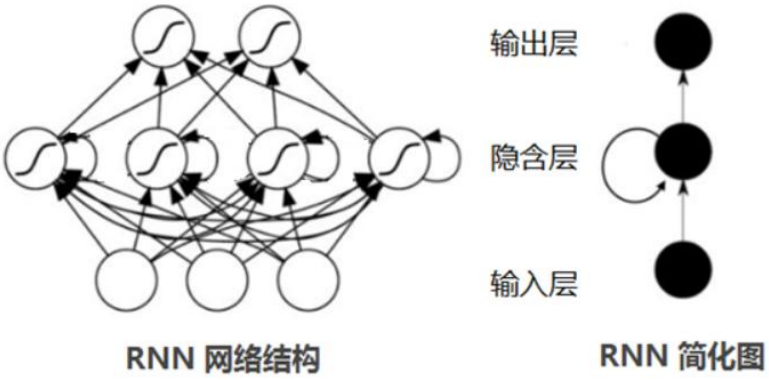
CNN 算法的优点包括模型的准确性高、对于图像处理有天然的优势、具有较强的特征抽象和泛化能力、可以降低算法的复杂度、并且可以利用 GPU 等硬件提高计算速度。

## (五)什么是 RNN 模型？

参考答案：

RNN(Recurrent Neural Network)，中文称作循环神经网络，它一般以序列数据为输入，通过网络内部的结构设计有效捕捉序列之间的关系特征，一般也是以序列形式进行输出。





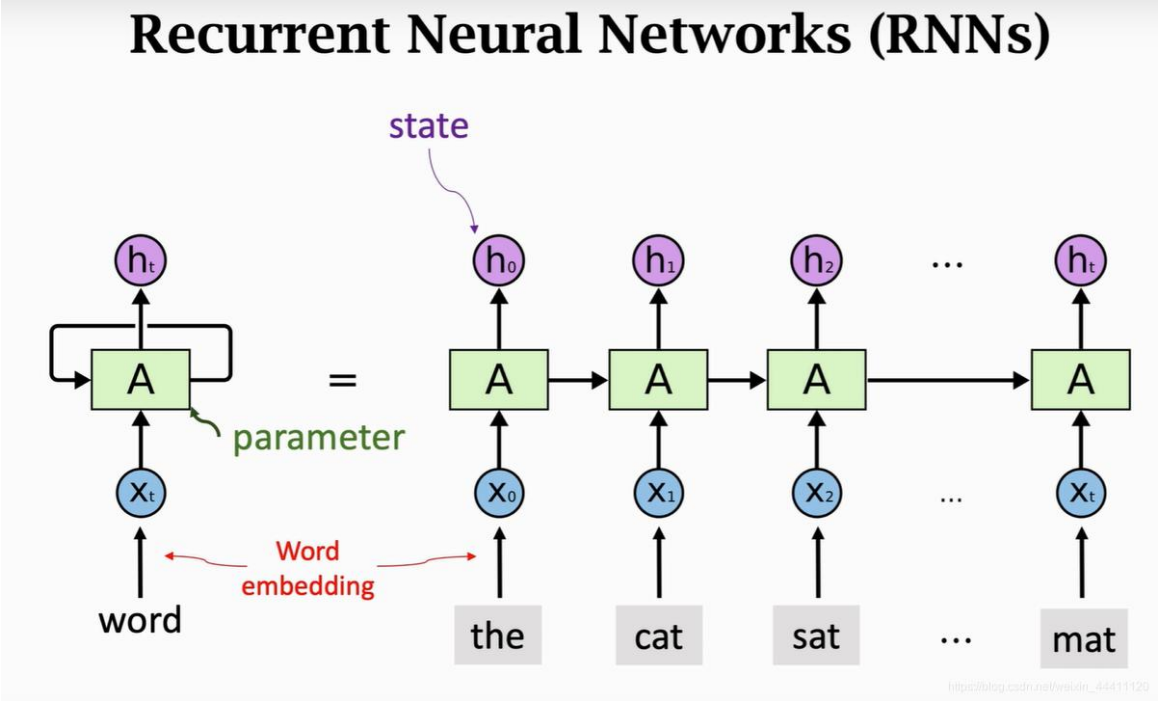

## (六)RNN 模型实现原理?

参考答案:

输入层：RNN 能够接受一个输入序列（例如文字、股票价格、语音信号等）并将其传递到隐藏层。

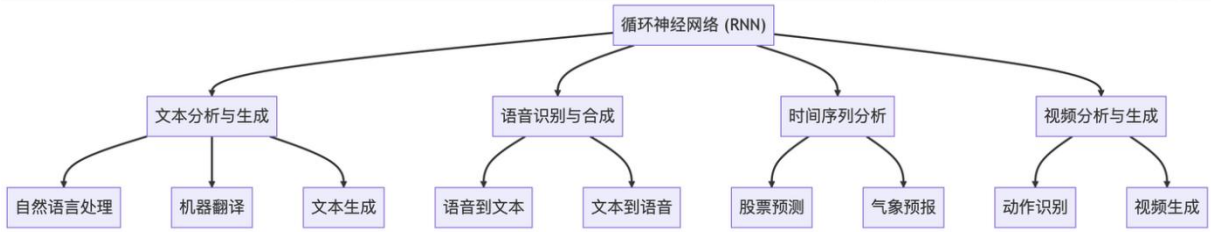
隐藏层：隐藏层之间存在循环连接，使得网络能够维护一个“记忆”状态，这一状态包含了过去的信息。这使得 RNN 能够理解序列中的上下文信息。

输出层：RNN 可以有一个或多个输出，例如在序列生成任务中，每个时间步都会有一个输出。




## (七)RNN 模型应用场景?

参考答案:




文本分析与生成

**1.自然语言处理**

RNN 可用于词性标注、命名实体识别、句子解析等任务。通过捕获文本中的上下文关系，RNN 能够理解并处理语言的复杂结构。

**2. 机器翻译**

RNN 能够理解和生成不同语言的句子结构，使其在机器翻译方面特别有效。

**3. 文本生成**

利用 RNN 进行文本生成，如生成诗歌、故事等，实现了机器的创造性写作。

语音识别与合成

**4. 语音到文本**

RNN 可以用于将语音信号转换为文字，即语音识别（Speech to Text），理解声音中的时序依赖关系。

**5. 文本到语音**

RNN 也用于文本到语音（Text to Speech）的转换，生成流畅自然的语音。

时间序列分析

**6. 股票预测**

通过分析历史股票价格和交易量等数据的时间序列，RNN 可以用于预测未来的股票走势。

**7. 气象预报**

RNN 通过分析气象数据的时间序列，可以预测未来的天气情况。

视频分析与生成

**8. 动作识别**

RNN 能够分析视频中的时序信息，用于识别人物动作和行为模式等。

**9. 视频生成**

RNN 还可以用于视频内容的生成，如生成具有连续逻辑的动画片段。

## (八)RNN 模型的优缺点？

参考答案：

优点：

能够处理不同长度的序列数据。

能够捕捉序列中的时间依赖关系。

缺点：

对长序列的记忆能力较弱，可能出现梯度消失或梯度爆炸问题。

训练可能相对复杂和时间消耗大。

## (九)什么是 GAN 模型?

参考答案:

GAN 由两个主要组成部分构成: 生成器 (Generator) 和判别器 (Discriminator) 。这两个部分通过对抗学习的方式相互竞争, 从而使得生成器能够不断提高生成逼真样本的能力，而判别器则不断提高辨别真伪样本的能力。

## (十)GAN 模型实现原理?

参考答案:

1.生成器 (Generator)

生成器的主要任务是接收一个随机噪声向量作为输入, 并将其转化为与真实数据相似的样本。生成器最初的输出可能非常随机，但随着训练的进行，它会逐渐生成更加逼真的样本。生成器的训练目标是欺骗判别器，使其无法准确区分生成的样本和真实数据。

2.判别器 (Discriminator)

判别器是一个二分类器，用于评估输入样本的真实性。它接收来自生成器的样本和真实数据，并尝试将它们正确分类为“真”或“假”。判别器的训练目标是尽可能准确地区分生成的样本和真实数据，使得生成器的输出更加逼真。

3.对抗学习过程

在训练过程中，生成器和判别器通过对抗学习相互博弈。生成器的目标是欺骗判别器，使其无法区分生成的样本和真实数据。而判别器的目标是尽可能准确地判断输入样本的真实性。这种对抗学习的过程持续进行，直到生成器生成的样本足够逼真，判别器无法有效区分真假为止。

## (十一)GAN 模型应用场景?

参考答案:

1.图像合成与编辑: GAN 可以用于生成高分辨率图像, 风格转换和图像增强。它也可以用于编辑图像，如将黑白照片转换为彩色。

2.视频生成：GAN 可以生成逼真的视频帧，扩展视频长度，甚至用于视频修复和插值。

3.自然语言处理：GAN 可以用于生成文本段落、对话和语音合成，提高机器翻译的质量等。

4.医学图像处理：GAN 可以用于生成医学图像数据，辅助医生进行诊断和手术规划。

5.游戏与虚拟现实：GAN 可用于创建虚拟角色、场景和游戏内容，提升游戏体验。

# 六、大数据模型类

## (一)什么是大模型?

参考答案:

我这边从 AI 产品经理的角度来看，首先 LLM 本身一种具有强大 NLP 能力的模型，通过 DL(深度学习)的方法和大量训练数据来生成，能够捕捉丰富的语言信息和深层的语义关联。实际上 LLM 在很大程度上重塑了 NLP 方向的研究和应用格局。通过更为先进的预训练和微调策略，让现在的大模型能够迅速适应多种

下游任务，显著降低了模型开发的难度和成本，另外 LLM 具有强大的生成能力，能够生成富有创意和质量较高的文本内容。

## (二)什么是 ROC 曲线?

参考答案:

ROC 曲线是接收者操作特征曲线的缩写，是一种用于评估分类模型性能的工具。ROC 曲线图像的横轴代表假阳性率，纵轴代表真阳性率。在 ROC 曲线图像中，对角线代表随机猜测模型的预测表现。ROC 曲线越靠近左上角，分类模型的性能越好。

下图中的蓝色曲线就是 ROC 曲线，它常被用来评价二值分类器的优劣，即评估模型预测的准确度。

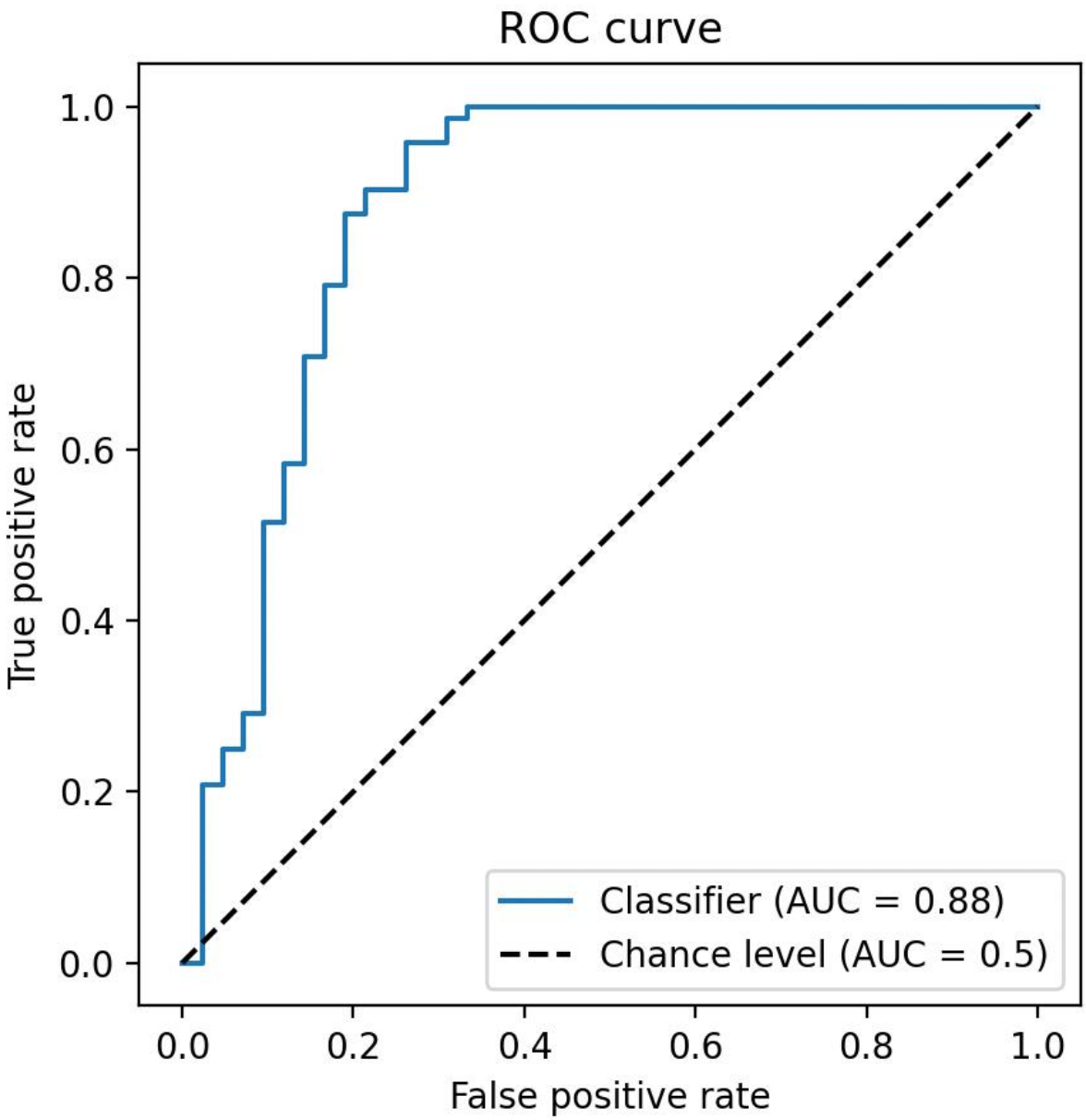
二值分类器，就是字面意思它会将数据分成两个类别(正/负样本)。例如：预测银行用户是否会违约、内容分为违规和不违规，以及广告过滤、图片分类等场景。篇幅关系这里不做多分类 ROC 的讲解。

坐标系中纵轴为 TPR（真阳率/命中率/召回率）最大值为 1，横轴为 FPR（假阳率/误判率）最大值为 1，虚线为基准线（最低标准），蓝色的曲线就是 ROC 曲线。其中 ROC 曲线距离基准线越远，则说明该模型的预测效果越好。(TPR: True positive rate; FPR: False positive rate)

ROC 曲线接近左上角：模型预测准确率很高

ROC 曲线略高于基准线：模型预测准确率一般

ROC 低于基准线：模型未达到最低标准，无法使用




## (三)什么是 AUC?

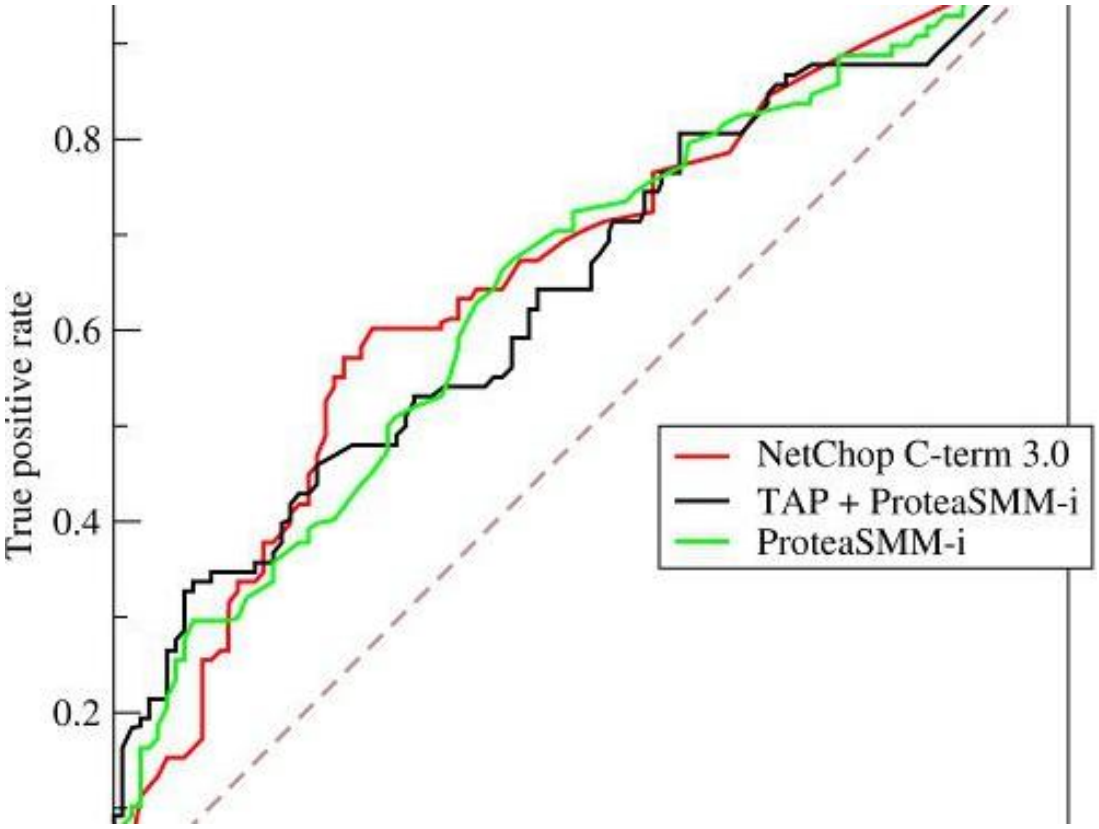
参考答案:

AUC 被定义为 ROC 曲线下的面积。往往使用 AUC 值作为模型的评价标准是因为很多时候 ROC 曲线并不能清晰的说明哪个分类器的效果更好，而作为一个数值，对应 AUC 更大的分类器效果更好。

其中，ROC 曲线全称为受试者工作特征曲线 ，它是根据一系列不同的二分类方式，以真阳性率感为纵坐标，假阳性率为横坐标绘制的曲线。

1.AUC 就是衡量学习器优劣的一种性能指标。从定义可知，AUC 可通过对 ROC 曲线下各部分的面积求和而得。

2.AUC 面积的意义：AUC 是衡量二分类模型优劣的一种评价指标，表示预测的正例排在负例前面的概率。




## (四)什么是 Transformer 模型?

参考答案:

Transformer 模型是由谷歌公司提出的一种基于自注意力机制的神经网络模型，用于处理序列数据。相比于传统的循环神经网络模型，Transformer 模型具有更好的并行性能和更短的训练时间，因此在自然语言处理领域中得到了广泛应用。

在自然语言处理中，序列数据的输入包括一系列文本、语音信号、图像或视频等。传统的循环神经网络（RNN）模型已经在这些任务中取得了很好的效果，

但是该模型存在着两个主要问题：一是难以并行计算，二是难以捕捉长距离依赖关系。为了解决这些问题，Transformer 模型应运而生。

作为一种基于自注意力机制的神经网络模型，Transformer 模型能够对序列中的每个元素进行全局建模，并在各个元素之间建立联系。与循环神经网络模型相比，Transformer 模型具有更好的并行性能和更短的训练时间。

Transformer 模型中包含了多层 encoder 和 decoder，每一层都由多个注意力机制模块和前馈神经网络模块组成。encoder 用于将输入序列编码成一个高维特征向量表示，decoder 则用于将该向量表示解码成目标序列。在 Transformer 模型中，还使用了残差连接和层归一化等技术来加速模型收敛和提高模型性能。

Transformer 模型的核心是自注意力机制（Self-Attention Mechanism），其作用是为每个输入序列中的每个位置分配一个权重，然后将这些加权的位置向量作为输出。

自注意力机制的计算过程包括三个步骤：

计算注意力权重：计算每个位置与其他位置之间的注意力权重，即每个位置对其他位置的重要性。

计算加权和：将每个位置向量与注意力权重相乘，然后将它们相加，得到加权和向量。

线性变换：对加权和向量进行线性变换，得到最终的输出向量。

通过不断堆叠多个自注意力层和前馈神经网络层，可以构建出 Transformer 模型。

对于 Transformer 模型的训练，通常采用无监督的方式进行预训练，然后再进行有监督的微调。在预训练过程中，通常采用自编码器或者掩码语言模型等方式进行训练，目标是学习输入序列的表示。在微调过程中，通常采用有监督的方

式进行训练，例如在机器翻译任务中，使用平行语料进行训练，目标是学习将输入序列映射到目标序列的映射关系。

## (五)什么是 ChatGPT 模型?

参考答案:

GPT 是 OpenAI 公司基于谷歌的 Transformer 语言模型框架而开发出来的技术。

GPT，英文全称是 Generative Pre-trained Transformer ，直译过来是生成型预训练-变换器 。名字前面加上 chat，即“聊天生成型预训练-变换器”。

从算法模式的版本上，ChatGPT 经历了 GPT-1（2018 年）、GPT-2（2019 年）、GPT-3（2020 年）和 InstructGPT（2022 年初）四个版本，未来有望生成新版本即 GPT-4（预计 2023 年）。

GPT-1（2018 年）：仅需要对预训练的语言模型做很小的结构改变，即加一层线性层，即可方面地应用于下游各种任务。

GPT-2（2019 年）：使用 zero-shot 设定，基本实现一劳永逸，训练一个模型，在多个任务上都能使用。

GPT-3（2020 年）不通过任何样例学习，而是利用少量样本去学习，更接近人脑学习模式。

InstructGPT（2022 年初）：经过多任务的微调后，能在其他任务上实现 zero-shot 预测，泛化能力极大提升。（InstuctGPT 可以理解成是 GPT-3 的微调版本，与 GPT-3 相比更擅长遵循指令，回答更真实，且有害情绪输出大幅下降）。

ChatGPT 可以理解成是 GPT-3.5 的微调版本，未来有望生成新版本即 GPT-4，相较于 InstructGPT，ChatGPT 效果更加真实，模型的无害性实现些许提升，编码能力更强。ChatGPT 使用的新的 AI 训练方法，加大“人”的反馈权重，进行训

练监督策略模型、训练奖励模型（Reward Mode，RM）、采用 PPO（Proximal Policy Optimization，近端策略优化）三个阶段的训练，在持续参数迭代的过程中，输入奖励模型，得到优化参数。且会不断重复第二和第三阶段，通过迭代，训练出更高质量的 ChatGPT 模型。

### (六)什么是 Diffusion 模型?

参考答案:

Diffusion 模型是一种深度生成模型，属于无监督学习中的概率模型，主要用于图像生成和视频预测等领域。

Diffusion 模型的工作原理是通过一系列高斯噪声逐步加入到原始图像中，直到图像变成纯高斯噪声。然后，模型通过去除噪声来还原图像。Diffusion 模型的特点是加入噪声的过程是可逆的，即噪声可以由原始图像逐步还原。因此，Diffusion 模型可以由给定的噪声图像还原出原始图像。

Diffusion 模型在图像生成方面具有很好的效果，可以生成高质量的图像。此外，Diffusion 模型还可以用于视频预测，即根据给定的前几帧预测未来帧。

## 七、技术基础类

### (一)什么是特征清洗、数据交换?

参考答案:

特征清洗（Feature Cleaning）是指在数据分析和机器学习任务中，对原始数据集中的特征进行处理和筛选，以提高模型的性能和准确性。特征清洗的目标是去除冗余、不相关或低质量的特征，同时保留与目标变量相关且有用的特征。

特征清洗可以包括以下几个方面的处理：

缺失值处理：对于含有缺失值的特征，可以选择删除该特征或使用合适的方法填充缺失值，如均值、中位数或众数等。

异常值处理：对于含有异常值的特征，可以通过设定阈值或使用统计方法来检测和处理异常值，如删除、替换或插值等。

数据类型转换：将特征的数据类型转换为适合分析和建模的形式，如将文本型特征转换为数值型特征，或将分类变量进行独热编码等。

特征选择：根据特征与目标变量之间的相关性或重要性，选择最具有预测能力的特征，可以使用统计方法、特征重要性评估或正则化方法等进行选择。

数据交换（Data Exchange）是指在不同系统或平台之间传输和共享数据的过程。在现实应用中，不同系统之间可能存在着数据格式、结构和接口的差异，因此需要进行数据交换来实现数据的互通。

数据交换可以采用多种方式，如文件传输、数据库连接、API 调用等。常见的数据交换格式包括 CSV（逗号分隔值）、JSON（JavaScript 对象表示法）、XML（可扩展标记语言）等。在数据交换过程中，需要确保数据的完整性、准确性和安全性，同时考虑数据量、传输速度和系统兼容性等因素。

数据交换在数据集成、业务合作和信息共享等场景中起到了重要的作用，能够促进不同系统之间的数据流动和协作，提高数据的利用价值和效率。

## (二)什么是过拟合和欠拟合?

参考答案：

过拟合（Overfitting）和欠拟合（Underfitting）是机器学习中常见的两个问题，涉及到模型在训练数据上的表现与在新数据上的泛化能力之间的平衡。

过拟合指的是模型在训练数据上表现良好，但在新数据上的预测能力较差。过拟合通常发生在模型过于复杂或训练数据过少的情况下。当模型过度拟合训练数据时，它会过分关注数据中的噪声和异常值，导致对新数据的泛化能力下降。过拟合的特征包括训练集上的误差很低，但验证集或测试集上的误差较高。

解决过拟合的方法包括：

增加训练数据量：通过增加更多的训练样本，可以减少模型对于训练数据的过度拟合。

减少模型复杂度：简化模型结构，如减少模型的参数数量、降低多项式次数等，以避免模型过于复杂而导致过拟合。

正则化（Regularization）：通过添加正则化项来限制模型的复杂度，例如 L1 正则化（Lasso）和 L2 正则化（Ridge），以减少模型对训练数据的过度拟合。

交叉验证（Cross-validation）：使用交叉验证来评估模型的泛化能力，通过将数据集划分为训练集和验证集，并多次进行训练和验证，以选择最佳的模型参数。

欠拟合指的是模型无法很好地拟合训练数据，导致在训练数据和新数据上都表现较差。欠拟合通常发生在模型过于简单或训练数据不足的情况下。当模型欠拟合时，它不能捕捉到数据中的复杂关系和模式，导致预测能力较弱。欠拟合的特征包括训练集和验证集上的误差都较高。

解决欠拟合的方法包括：

增加模型复杂度：增加模型的参数数量、引入更多的特征等，以提高模型的灵活性和拟合能力。

改进特征工程：对原始数据进行更好的特征提取和选择，以提供更有信息量的特征。

增加训练数据量：通过增加更多的训练样本，可以提供更多的信息来改善模型的拟合能力。

调整模型超参数：调整模型的超参数，如学习率、正则化参数等，以找到更好的模型配置。

过拟合和欠拟合是机器学习中需要关注和解决的问题，通过合适的方法和技术，可以使模型在训练数据和新数据上都能够取得良好的表现。

## (三)什么是跨时间测试和回溯测试?

参考答案:

跨时间测试也叫 OOT 测试，是测量模型在时间上的稳定性。回溯测试是用过去- -段时间的真实数据构造出一个模拟的环境(回溯环境)，让模型在历史的那段环境中运行，得到历史某个时间点的模型结果。

一般来说，跨时间测试是在模型上线之前就应该要做的事情。回溯测试是指模型已经存在并已经上线了，想要看模型在历史某个时间点的数据表现时候，进行的测试。

## (四)什么是训练集、验证集和测试集?

参考答案:

训练集(training set)

顾名思义指的是用于训练的样本集合,主要用来训练神经网络中的参数.

验证集(development set 或 validation set)

用于验证模型性能的样本集合.不同神经网络在训练集上训练结束后,通过验证集来比较判断各个模型的性能，有时候也被称为开发集。

测试集(test set)

对于训练完成的神经网络,测试集用于客观的评价神经网络的性能.

## (五)你之前负责产品中使用的最核心的算法是什么？这种算法有哪些优缺点？

参考答案:

第四章有详细介绍

## (六)你对深度度学习有哪些了解？深度学习的应用场景有哪些？

参考答案:

深度学习是一种人工神经网络的应用，是机器学习的分支之一。它是通过构建多层神经网络来模拟人类的神经系统，从而实现对大量数据的自动分类和预测。

深度学习的最大特点是通过多层次的特征提取和组合来实现高效的数据处理。深度学习的基本原理是通过前向传播算法，将输入的数据通过多层神经网络，一层一层地进行特征提取和组合，最终得出分类或预测结果。

**应用场景**

深度学习的应用领域非常广泛，包括自然语言处理、图像识别、语音识别、智能推荐等。

其中，在图像识别领域，深度学习已经取得了非常显著的成果，例如在 ImageNet 大规模视觉识别挑战赛中，深度学习的表现已经超过了人类的识别能力。

例如在医疗领域中，深度学习已经开始被用于医学图像的分析和疾病预测等方面。

在智能推荐领域，深度学习也被广泛应用于产品推荐和广告投放等方面。

## (七)什么是机器学习?

参考答案:

机器学习（Machine Learning）是对研究问题进行模型假设，利用计算机从训练数据中学习得到模型参数，并最终对数据进行预测和分析的一门学科。

## (八)机器学习的应用场景都有哪些?

参考答案:

图像和语音识别：机器学习可以通过训练模型来实现图像和语音识别，如人脸识别、语音识别和文字识别等。

自然语言处理：机器学习可以用于文本分析、情感分析、机器翻译、问答系统等自然语言处理任务。

推荐系统：机器学习可以通过学习用户的行为和兴趣来实现个性化推荐，如电商网站的商品推荐、音乐推荐等。

金融风控：机器学习可以用于金融领域的风险控制和欺诈检测，如信用评估、反洗钱等。

医疗健康：机器学习可以应用于医疗领域，如疾病诊断、医学影像分析、药物研发等。

自动驾驶：机器学习可以应用于自动驾驶技术，如图像识别和预测等。

工业生产：机器学习可以应用于工业领域的自动化生产和设备维护，如设备故障预测和优化生产等。

总之，机器学习具有广泛的应用领域，可以用于图像和语音识别、自然语言处理、推荐系统、金融风控、医疗健康、自动驾驶、工业生产等方面。通过机器学习，可以实现更智能、更高效、更准确的决策和服务，为人类带来更多的便利和效益。

## (八)逻辑回归相比于线性回归，有什么区别?

参考答案:

区别：性质不同、任务定位不同、输出值不同、损失函数不同等。

**性质不同**

逻辑回归是一种广义的线性回归分析模型；线性回归是利用数理统计中回归分析，来确定两种或两种以上变量间相互依赖的定量关系的一种统计分析方法。

逻辑回归常用于数据挖掘，疾病自动诊断，经济预测等领域；线性回归常运用于数学、金融、趋势线、经济学等领域。

**任务定位**

线性回归用于回归任务；逻辑回归用于分类任务。

**输出值**

线性回归输出连续值；逻辑回归输出概率值；本质是因为逻辑回归使用了 sigmod 函数进行了映射 ，将值域映射到(0,1)，在二类任务中，若大于 0.5，则为某个类，小于 0.5，为另一类。

**损失函数**

线性回归采用 MSE 损失函数，逻辑回归采用交叉熵损失函数。

**线性回归**

在 LR 中，将线性回归的结果通过 sigmod 函数映射到 0 到 1 之间，映射的结果刚好可以看做是数据样本点属于某一类的概率，如果结果越接近 0 或者 1，说明分类结果的可信度越高。这样做不仅应用了线性回归的优势来完成分类任务，而且分类的结果是 0~1 之间的概率，可以据此对数据分类的结果进行打分。对于线性不可分的数据，可以对非线性函数进行线性加权，得到一个不是超平面的分割面。

**逻辑回归**

逻辑回归虽然叫做回归，但是其主要解决分类问题。可用于二分类，也可以用于多分类问题。由于线性回归其预测值为连续变量，其预测值在整个实数域中。而对于预测变量 y 为离散值时候，可以用逻辑回归算法（Logistic Regression）逻辑回归的本质是将线性回归进行一个变换，该模型的输出变量范围始终在 0 和 1 之间。

### (九)你能介绍一下 KNN/朴素贝叶斯/SVM/CNN/Diffusion/NLP 的原理吗？你熟悉哪几种深度学习和机器学习算法？都有哪些区别

参考答案：

第四章有详细介绍

## 八、工作场景类

### (一)AI 算法工程师说你的需求实现不了怎么办？

参考答案：

从产品工作一开始我们就有必要执行三部曲：确立专家效应→找到根本原因→使用沟通软技能。

### (二)工作中做的最失败的事情/项目/遇到的最大困难是什么？

参考答案：

**问题原因：**

然而，项目在开始阶段就遇到了困难。首先，需求分析不足，导致我们没有完全了解市场需求和竞争情况。其次，团队内部的沟通问题也显现出来，不同部门之间的协作不够紧密。最重要的是，项目管理方面存在问题，没有明确定义的里程

碑和时间表，导致进度延误。

**采取的行动：**

面对这个局面，我作为项目团队的一员，采取了一系列行动来纠正问题。首先，我促使团队进行了深入的市场研究，以更好地理解客户需求和竞争对手的情况。其次，我倡导加强跨部门协作，促进信息共享和合作。最后，我与项目管理团队合作，制定了明确的项目计划，并设定了可行的里程碑和截止日期。

**教训和启示：**

尽管我们采取了这些行动，但最终项目仍然失败了。这次经历让我深刻认识到了几个重要的教训。首先，市场研究和需求分析是项目成功的关键，不能忽视。其次，有效的团队协作和沟通对于项目的顺利进行至关重要。最后，良好的项目管理和计划是确保项目按时交付的关键因素。

## (三)请说说你们产品的主要竞品是谁?

参考答案：

针对竞品公司，我需要了解具体的市场情况和行业背景。在人工智能行业中，有许多具有竞争力的公司。以图像识别技术为例，商汤和依图是知名的 AI 技术公司，它们在图像识别领域有着深厚的技术积累和丰富的应用经验。

## (四)如果公司研发资源不足以实现你想要的功能？怎么办?

参考答案：

**确保你的方案是完整且精简的闭环**

一个磕磕绊绊的方案落地一定会受阻，同时也会降低你的信誉度，让你在之后方案的推进中举步维艰。

**拆分阶段**

当资源固定且你需要分配出的任务复杂度偏高时，将任务拆分成多个阶段，分步骤出成果的是更好的方案，快速出成绩能保证你的“大棋”不会死在棋局成型前。

**尽量评估精准**

规避开发因为对任务认知不够，评估出的时间超出你的容错范围；如果能力不足无法实际评估，第一时间寻求对方上级的支持。

**没人愿意做对自己毫无意义的事情**

在沟通前搞清对方的 okr，找到可以和你的需求和他的贴合点，会让沟通结果达到双方满意的程度。

**施加压力**

每个人在工作中都会存在压力，那么你也可以作为施压方，你可以认为这是在 CPU 他人，但屠龙者终成恶龙只是时间问题，最重要的如何善用手头的一切资源。这里要注意的点是对事不对人，勿带情绪做事，情绪化对于成年人来说是一件特别幼稚的事。

当然，也有可能通过各种方式的尝试后，还是没有得到理想结果；如果直接搁置最终结果是不利于自己的成果产出的，该怎么破局呢？可否在市面上找到一些类似的模版或无代码开发平台，自力更生呢？

## (五)训练模型时数据集都有哪些来源？

参考答案：

机器学习用于模型训练的数据来源主要包括公开数据集、自有数据集、合作伙伴提供的数据集等。其中公开数据集是比较常见和容易获取的，例如 mageNet、COCO、MNIST 等，自有数据集则是企业或组织自己收集、整理和标注的数据集，合作伙伴提供的数据集则是通过与其他企业、组织或个人的合作来获得的数据集。

## (六)工作中用什么样的方法清洗数据?

参考答案:

数据清洗的主要包括：纠正错误、删除重复项、统一规格、修正逻辑、转换构造、数据压缩、补足残缺/空值、丢弃数据/变量。

## (七)模型构建流程通常包括几个阶段?

参考答案:

1.筹备阶段：在这个阶段，我们与团队成员和相关利益相关者合作，明确项目的目标和范围。我们确定项目所需的数据集、技术要求和资源配备，并制定项目计划和时间表。

2.数据准备阶段：在这个阶段，我们收集和准备所需的数据集。这可能涉及数据的获取、清洗、标注和格式转换等工作。我们还会对数据进行分析和探索，以确保其质量和适用性。

3.模型开发阶段：在这个阶段，我们根据项目的需求和目标选择适当的机器学习算法和模型架构。我们进行特征工程，选择和提取最相关的特征，并进行模型训练和调优。在这个阶段，我们可能会进行多次迭代，以改进模型的性能和准确性。

4.模型评估阶段：在这个阶段，我们对训练好的模型进行评估和验证。我们使用验证数据集对模型进行测试，评估其性能和泛化能力。如果需要，我们可能会进行模型调整和改进，以达到预期的效果。

5.上线部署阶段：在这个阶段，我们将训练好的模型部署到实际应用环境中。我们将模型集成到相应的系统中，并进行系统测试和性能优化。同时，我们确保模型的稳定性、安全性和可扩展性。

# 九、行业认知类

## (一)你怎么看待 AI 或者人工智能行业？对于整个 AI 行业有哪些认知?

参考答案:

我对 AI 或人工智能行业持乐观态度。目前，人工智能已经在许多行业得到广泛应用，如金融、医疗、零售和制造等。AI 技术的不断进步和成熟，为企业提供了巨大的机会和挑战。我认为人工智能行业正在迅速发展，并且将持续成为未来的关键技术领域。

我密切关注 AI 行业中的一些新技术和新应用。例如，近年来，深度学习、自然语言处理和计算机视觉等技术的突破，使得人工智能在图像识别、语音识别、智能助理等领域取得了重大进展。此外，强化学习和自动驾驶等领域也受到了广泛的关注和投资。

**认知**

我的独特观点是，AI 不仅仅是一项技术，而是一种推动社会进步和创新的力量。随着技术的不断发展和应用的扩大，AI 将深刻改变人们的工作方式、生活方式和社会结构。同时，我认为在 AI 发展的过程中，我们需要关注伦理、隐私和安全等重要问题，以确保人工智能的发展能够造福整个社会。

## (二)结合我们公司的业务场景？通过 AI 技术可以做哪些工作来提升用户体验?

参考答案:

通过 AI 技术，可以在业务场景中提升用户体验的多个方面。举例来说，对于电商平台，AI 可以通过推荐系统个性化推荐商品，提供更准确的搜索结果，

从而提高用户的购物体验。在客服领域，AI 可以通过自然语言处理和智能对话系统，实现智能客服，提供快速、准确的问题解答和服务。在智能家居领域，AI 可以实现智能语音助手，使用户可以通过语音控制设备，提供更便捷的生活体验。总的来说，AI 技术可以帮助优化业务流程，提高效率，减少人为错误，提供个性化的服务，从而提升用户的满意度和体验。