《人工智能与商业应用》作业

第一部分：个人作业

一、引言

开源模式已成为推动人工智能技术快速迭代与普惠化的核心力量，从代码托管平台到模型开源社区，从基础算法框架到垂直领域应用，开源生态深刻影响着AI技术的研发路径、产业协作模式与商业化逻辑。本次作业要求同学们围绕AI领域的开源主题展开调研，系统梳理开源的核心内涵与商业运作模式，通过分析典型平台与案例，深入理解开源在AI技术产业发展中的关键作用，培养对技术生态与商业逻辑的综合分析能力。

二、作业要求：**本次作业为个人独立作业，无需组队，每位同学需独立完成调研、分析与报告撰写，确保内容原创性。**

请围绕以下三个维度展开调研，形成结构化分析内容：

1. 开源的核心定义与AI开源平台实践，开源是什么？

明确开源的官方定义，并结合诸如GitHub的代码托管场景、Hugging Face的大模型开源社区场景等平台为典型案例，分析平台的核心功能、用户群体，以及它们在AI算法共享、模型迭代、工具协作中的具体作用。

2. 开源对人工智能发展的核心价值,为什么要开源？

开源对于人工智能发展有着诸多价值和意义，其关键驱动作用可从多维度展开。在**推动技术创新**方面，它能降低研发门槛、加速技术迭代；在**优化产业生态**方面，它助力中小企业技术落地、促进跨企业协作；在**促进资源普惠**方面，它支撑高校科研开展、方便个人开发者学习。请同学们结合具体实例，阐述为何开源能够成为AI领域发展的关键驱动力？

3. AI开源项目的主流商业模式，开源的商业模式是怎样的？

调研当前开源领域可持续运营的典型商业模式，有开源+企业服务+付费增值功能+硬件配套等，说明你最喜欢的商业模式以及运作逻辑与适用场景。

三、作业提交说明

1. **篇幅限制：**报告整体字数需合理控制，排版后不超过**2页**（建议采用“**宋体、小四号字体、1.5倍行距**”格式，图表需计入篇幅）。

2. **内容规范：**需逻辑清晰、重点突出，可适当使用表格、示意图辅助分析；所有引用的资料需标注明确来源；提交前需自行核查内容真实性，若发现借助工具生成无事实依据的表述，将影响作业得分。

第二部分：团队作业

一、引言

我们本次作业需要完成基于借贷数据的违约预测模型，请利用给定的借款与还款数据（LC 表、LP 表与 LCIS 表），设计并实现一个违约预测模型，判断一个借款人是否会在贷款周期内发生逾期。

二、作业要求

本次作业包括以下 4 部分：

1. 团队建立和分工

本次作业**三人为一组**，共同完成。请同学们自行结组，请在结组后将名单填至腾讯文档

https://docs.qq.com/sheet/DV2Zmd0R6S3dWeUJs?tab=BB08J2

2. 作业提示

1) 数据理解

首先需要描述 LC、LP、LCIS 表的结构与含义。接着确定预测目标，其中目标变量（label）为是否逾期，该变量从 LP 表生成，若还款状态含有“逾期”则为 1，否则为 0。特征包括借款金额、期限、利率、初始评级、认证信息、历史借款行为等。这部分需要同学将业务逻辑梳理成技术语言进行描述，体现在报告中。

2) 数据预处理

在数据预处理阶段，对于缺失值要进行填补、删除或者插值处理；针对异常值，像年龄过大或过小、借款金额极端值等情况也要进行处理。同时需要进行编码转换，将类别变量（如性别、认证、借款类型、评级）转换为 one-hot 编码，对日期类变量（如借款成功日期、还款日期）提取年、月、是否周末等信息。

3) 特征工程

此阶段要构造新特征，例如借款人历史逾期率，其计算方式为历史逾期还款期数除以历史成功借款次数。借款人还款能力指标，即历史正常还款期数除以

（历史逾期+正常还款期数）。还可以用借款金额除以收入 proxy（若缺数据可 用“评级+认证”特征近似）。之后进行特征选择，可采用相关系数、卡方检验、模型特征重要性等方法。

4) 模型训练

选择至少两类模型(模型可以查阅文献/AI 工具/开源社区 Git 等，需要在报告中做引用标注)，其中包括作为基线模型的逻辑回归，以及决策树、随机森林、XGBoost等模型。采用训练集/验证集/测试集划分（比例为 70%-15%-15%），并进行超参数调优，例如使用Grid Search CV。

5）模型评估

使用分类指标来评估模型，如Accuracy、Precision、Recall、F1-score、ROC–AUC 等，根据自己所掌握知识和对项目理解灵活选择。同时绘制 ROC 曲线和混淆矩阵，对比不同模型的效果并解释原因。

6）结果与讨论

至少选择两个模型进行对比，要明确哪个模型效果最好，找出哪些特征对违约预测贡献最大（例如初始评级、历史逾期次数），还要探讨模型可能的改进方向，如使用深度学习、引入时间序列特征等。

3. 项目开发

以小组为单位进行项目开发，团队成员根据项目内容和职位要求，通过沟通、开发和测试完成该作业，尝试通过不断迭代上述过程提升项目质量，进度较快的小组可在完成作业基本要求的基础上讨论其他有效模型写进报告（小组加分）。

4. 项目总结

在完成项目的开发工作后，团队成员对整个项目的开发过程、各自在项目中承担的工作进行总结，形成一份完整的项目总结报告。

三、项目团队和职位说明

每位同学担任研发团队中的一个职位，完成相应工作，并在该作业报告完成后针对自己的职位和工作过程进行总结，叙述自己与团队中与其他职位同学的沟通、合作过程和具体工作内容。下面为同学们列举了一些职位及要求作为参考，同学们根据自己的能力和团队的需要选择职位，也可以根据具体情况做出适当的调整，只要能够为团队做出贡献、参与团队研发并且完成对应工作即可：

1. 客户

1) 工作内容：为本次作业的项目寻找一个合适的行业背景和应用场景，该项目可以扩展到所有信贷领域；针对所选择的场景，提出具体的产品要求，与项目经理对接具体功能要求和指标；在完成目标后可以根据项目进度和作业完成时间进一步提升要求，例如模型效果等。
2) 报告内容：项目的行业背景和应用场景，项目内容、功能要求和具体指标要求，对研发团队最终模型的评价和总结等。
3) 能力要求：对行业背景和应用场景有所了解，能够根据项目内容提出具体的要求和指标。

2. 项目经理

1) 工作内容：收集需求、分析需求、工作协调、进度跟踪等，负责项目的质量、进度；负责与客户的沟通以及引导、协调研发团队内的研发人员开展工作；根据项目进度和完成时间对项目周期进行合理的安排和调整。
2) 报告内容：与客户的沟通过程和需求分析，项目进度总结，与研发团队内成员的沟通和协调过程总结，项目最终成果的总结和分析等。
3) 能力要求：熟悉机器学习和相关项目研发的全流程，具有较强的沟通、合作和协调能力，合理安排、掌握项目进度。

3. 数据科学家

1) 工作内容：理解客户业务需求，针对客户需求和数据集质量提出合适的数据处理方法（数据清洗与预处理）和深度学习模型架构。
2) 报告内容：对数据集相关内容的总结（例如数据集的规模、质量等），项目中所使用到的各种数据处理方法、使用理由以及应用效果等。
3) 能力要求：了解数据处理的意义、流程和基本方法，对数字图像处理和深度学习的基础知识有所了解更佳，具体代码实现可以通过与项目经理和机器学习算法工程师进行沟通、合作来完成。

4. 机器学习算法工程师

1) 工作内容：理解客户业务需求，与项目经理沟通具体要求，与研发团队内的其他成员共同合作，利用 Python、PyTorch 等工具完成模型的构建、训练和测试。
2) 报告内容：模型结构的介绍、使用理由及应用效果等。
3) 能力要求：了解数字图像处理、机器学习、深度学习的基本知识，对 Python 程序设计和深度学习框架有所了解更佳，根据作业教程和示例代码完成本次作业的程序设计内容。

5. 测试工程师

1) 工作内容：理解客户业务需求，根据客户描述，与项目经理和机器学习算法工程师沟通制定测试规范、测试方案和具体的测试指标；完成项目产品的测试和分析工作；参与测试结果的评审工作。
2) 报告内容：测试规范、测试方案和具体的测试指标的说明和总结；项目迭代过程中不同模型的对比和分析；最终产品性能的总结和分析等。
3) 能力要求：了解测试的规范和过程，根据作业教程学习如何运行代码来对模型进行测试和分析。

四、作业提交说明

本次作业要求提交的内容包括以下几个部分：

1. 清洗后的数据集

P2P_PPDAI_DATA 文件夹中存储了给本次作业提供的数据集，包含了成交时间从 2015 年 1 月 1 日到 2017 年 1 月 30 日的 328553 支信用标。该数据未经过任何筛选、处理和分割，请同学们自行对原始数据进行分析、处理、清洗，剔除掉不符合要求的数据，将清洗后的数据集打包成 zip 压缩文件提交。

2. 代码

请提交最终编写完整后的代码，保证能够正常运行，在批改作业时代码测试和作业报告均作为评分标准。如果需要提交多个版本，或对代码有其他说明，请在小组报告中注明。

3. 小组报告

对本次作业的完成过程和最终结果进行总结分析。具体参考内容如下，其中项目概述和项目总结必须包含在报告内容当中，其他部分同学们可以根据具体的分工情况和职位选择情况进行调整：

1) 项目概述（由团队成员共同撰写）：对本次作业中借贷数据的违约预测模型中所解决的问题和实现的效果进行综述，并对团队成员、负责职位和具体工作内容进行详细描述等。
2) 客户需求描述（由负责模拟客户的团队成员撰写）：描述项目的行业背景和应用场景，介绍项目内容、功能要求和具体指标要求，对研发团队最终模型进行评价和总结等。
3) 项目经理工作报告（由担任项目经理的团队成员撰写）：与客户的沟通过程和需求分析；项目进度总结；与研发团队内成员的沟通和协调过程总结；项目最终成果的总结和分析等。
4) 数据科学家工作报告（由担任数据科学家的团队成员撰写）：对原始数据的质量进行分析（原始数据文件夹 P2P_PPDAI_DATA，未经过任何筛选和处理，存在不可用数据（例如部分字段 NULL）；训练和测试时需要将原数据清洗，并总结不可用数据的类型；分析产生不可用数

据的原因；思考如何避免产生不可用数据；提出更加高效和有效的数据采集方式。

5) 机器学习算法工程师工作报告（由担任机器学习算法工程师的团队成员撰写）：介绍模型结构与功能，可以采用图示、表格、文字等方式进行描述，并说明该模型的选择过程及原因；描述超参数的调节过程及对应的训练效果变化过程；介绍数据预处理的具体实现方式等。

6) 测试工程师工作报告（由担任测试工程师的团队成员撰写）：描述测试规范、测试方案和具体的测试指标；对项目迭代过程中的不同模型进行对比和分析；通过混淆矩阵、输出的详细分类结果等对最终产品的性能进行总结和分析等。

7) 项目总结（由团队成员共同撰写）：针对本次作业中选择的行业背景和应用场景，分析领域内其他客户的更多需求，设想具体的商业模式，分析产品的推广方式，探讨产品投入实际应用所需达到的性能要求（如准确率）。

第三部分：作业提交要求

✓ 提交格式：

1) 个人作业：以 **PDF** 格式保存报告，命名方式为“**学号-姓名-个人作业**”。

2) 团队作业：以压缩包的方式提交作业，将编写完成后的 **代码、报告和清洗后的训练集** 添加至 **zip 压缩文件** 并命名为“**组号-组员1姓名-组员2姓名-组员3姓名.zip**”(示例“1-张三-李四-王五.zip”)，**每人都要将作业提交到系统中。**

✓ 提交渠道：将作业提交至 **交大CANVAS系统** 。

✓ 截止时间：**11 月 17 日 22:00**（逾期提交无效）。