孟晓

女 | 年龄：37岁 | 17628072710 | santochaoya@yahoo.com.hk
14年工作经验 | 求职意向：算法工程师 | 期望薪资：20-30K | 期望城市：成都

个人优势

具有 5 年以上数据科学、数据挖掘、数据建模、数据分析和生成式人工智能（GenAI），专注于 金融、通信和工程咨询 等领域。

在自然语言处理（NLP）及大型语言模型（LLM）积累了丰富的实践经验，熟练掌握 HuggingFace Transformers、LangChain、RAG 及 Azure OpenAI等技术。

同时专注于时间序列预测（Forecasting）与分类（Classification）模型，精通 LSTM、ARIMA 及多种机器学习与深度学习算法，并成功运用于多个行业场景，熟悉并实践完整端到端深度学习项目开发生命周期。

同时具备 MLOps、数据管道构建、API 开发与部署 经验，能够高效实现端到端的 AI 解决方案。

具备良好的沟通和跨团队协作能力，能够与各个团队紧密合作，确保交付既符合业务需求、可行性。能快速学习新技术、新业务，适应多样化和快速变化的项目需求。

工作经历

Epam System (Singapore) Data Scientist 2024.03-至今

- 利用各种时间序列技术、深度学习和机器学习模型，构建端到端金融工具的预测模型和时间序列预测。
- 开发用于新闻分析的 NLP 解决方案，包括基于 Transformer 的模型和 GenAI 处理多个 NLP 任务。
- 通过假设检验和时间序列方法，对市场进行统计分析、生存分析和 A/B 测试。
- 构建基于 WebSocket 和 RESTful API 的实时数据处理平台。
- 实现金融数据的端到端数据管道，包括网页爬取和数据库集成。
- 开发自动化建模和分析工具，以优化数据清洗、特征工程和模型评估的工作流程。

EPAM 亿磐系统科技（成都）有限公司 Data Scientist 2021.09-2024.03

内容：

- 利用各种时间序列技术、深度学习和机器学习模型，构建端到端金融工具的预测模型和时间序列预测。
- 开发用于新闻分析的 NLP 解决方案，包括基于 Transformer 的模型和 GenAI 处理多个 NLP 任务。
- 通过假设检验和时间序列方法，对市场进行统计分析、生存分析和 A/B 测试。
- 构建基于 WebSocket 和 RESTful API 的实时数据处理平台。
- 实现金融数据的端到端数据管道，包括网页爬取和数据库集成。
- 开发自动化建模和分析工具，以优化数据清洗、特征工程和模型评估的工作流程。

业绩：

客户高度认可：在项目中和客户合作愉快，凭借扎实的技术能力和高效的沟通，赢得了客户的信任和认可。客户希望能更紧密地合作，因此要求我 Relocate 到新加坡现场办公，参与关键项目的推进和实施。

亚信科技（成都）有限公司 算法工程师 2021.05-2021.09

● 负责构建和优化 机器学习模型，用于评估用户体验、优化客户满意度，并进行风险管理。
● 针对电信基站的时序数据，开发定制化的机器学习模型和多步算法，以提升预测精度和异常检测能力。
● 在多个项目中负责从数据处理、模型训练到优化部署的全流程开发，涵盖时序预测、异常检测、增量学习、用户满意度分析等领域

Mott MacDonald 机器学习 2019.05-2021.01

1. 数据挖掘及处理，搭建连接API、爬虫、AWS S3 云存储、处理数据云函数(A zureFunction、CI/CD)
2. 搭建GIS平台、建立时间序列模型、机器学习等进行预测及分析
3. 异常点处理、提取显著事件等 用PowerBI(+Python、R)进行数据可视化面板开发
4. 对已有模型、算法、代码进行优化
5. 开发Python工具优化自动化工作流程
6. 编写项目及模型文档

项目经历

实时 RFQ 分析与债券偏向评估平台 数据工程师、数据分析师 2024.03-2024.12

构建了一个实时 RFQ（询价单）分析平台，用于评估 债券 Skew 和对手方购买意愿，帮助交易员在谈判中优化定价策略。系统基于 WebSocket 进行实时数据流处理，并通过 Flask RESTful API 计算 债券库存、市场需求、历史成交数据 等关键指标，

● 实时数据处理：基于 WebSocket 低延迟处理 RFQ 订单流，实现实时数据展示和计算。
● 基于 MariaDB 设计多表查询，实现 Skew 计算引擎，根据 库存水平、成交量、市场波动 计算债券购买意愿评分，为交易员提供定价策略参考
● 采用 Flask 构建高并发 API，支持交易数据查询、库存分析、Skew 计算等功能。
● 开发 Auto-Profiling 工具，实现 EDA、特征工程和数据可视化 自动化。

技术栈：
WebSocket、MariaDB、Flask、OpenShift、Bitbucket、Python、Pandas、Scikit-learn、Matplotlib、Seaborn

金融新闻分析与实体提取系统 数据科学家 2022.09-2024.03

本项目旨在构建一个自动化金融新闻分析、实体提取、价格走势分析系统，从订阅的邮件账户中抓取新闻内容，LLM，结合 LangChain 动态融合外部数据，提高金融新闻摘要的准确性和上下文相关性。集成了 实体提取，实体匹配、主题建模、情感分析等 NLP 任务，项目进一步分析新闻情绪对债券价格波动的影响，结合历史债券数据与情感分析模型，探索市场情绪与债券价格 movement 之间的关联性。

● 开发爬取脚本，自动从订阅邮件账户中提取新闻文章。
● 构建端到端 RAG 管道，结合 LangChain 提升模型的金融新闻摘要能力，使其能够动态融合外部数据。
● 实现命名实体识别（NER），用于提取新闻中的金融机构、股票、公司名称等实体。
● 应用实体匹配技术，利用 BERT 模型 进行匹配，并将提取的实体与数据库记录关联。
● 开展主题建模、主题摘要、情感分析，结合 BERT 和 LLM 模型分析新闻内容。
● 分析新闻情绪对债券价格的影响，结合历史债券交易数据，研究市场情绪如何驱动债券价格 movement。
● 研究与优化 Prompt，评估和优化 NLP 处理管道，提升文本处理效果。
● 实现 LLM 流水线，对 NER、实体匹配、文本分类、情感分析 等 NLP 任务进行优化和微调。

技术栈：

NLP、HuggingFace Transformers、BERT 和 GPT 系列模型（BERT-NER、RoBERTa、FinBERT、DistilBERT、BERTopic、chat-bison、text-bison、claude、GPT、Gemini、Llama、Qwen）、Vertex AI、LangChain、ASC、Fuzzywuzzy、Beautiful Soup、Scrapy、PyTorch、Pandas、MariaDB

新闻发布与情绪对债券价格波动的统计分析　数据分析师　2023.10-2024.02

本项目分析新闻发布及市场情绪对债券价格变动的影响，采用 A/B 测试、假设检验、时间序列分析和生存分析 量化新闻对市场波动的作用，并建立回归模型评估情绪与价格变动的关系。

- A/B 测试：设计并执行新闻发布 vs. 非发布对债券价格影响的对比实验。
- 假设检验：分析不同情绪新闻对债券价格波动的影响显著性。
- 生存分析：评估新闻影响持续时间，研究市场反应的衰减趋势。
- 时间序列建模：采用 ARIMA 预测债券价格波动及新闻事件对市场的长期影响。
- 回归分析：量化新闻情绪与债券价格变化之间的关系，探索市场反应机制。

技术栈

Python、Pandas、Scikit-learn、Matplotlib、回归分析、时间序列分析、情感分析、生存分析

基于 LSTM 的债券价差预测与自动化部署系统　数据科学家，MLE　2021.12-2022.10

构建债券价差预测模型，使用 LSTM、XGBoost、LightGBM 进行实验，并通过 MLflow 进行版本管理，结合 CI/CD、Airflow、Jenkins 和 OpenShift 实现自动化训练、微调和部署，同时设计实时监控系统确保长期稳定性

- 数据处理：收集并清理债券 tick 数据和信用利差数据，完成 EDA 和特征工程。
- 模型开发：试验LSTM、XGBoost、LightGBM 预测债券价差，并利用 MLflow 进行版本管理、超参数优化及模型评估。
- 自动微调：使用 Airflow 进行每日自动微调，并结合 Evidently AI 进行数据漂移检测。
- 自动化部署：通过 Jenkins + OpenShift 实现模型自动更新、测试与生产部署。
- 实时监控：采用 Prometheus 监测推理结果，异常时触发数据漂移检测并生成报告。

技术栈：

LSTM、XGBoost、LightGBM、随机森林、回归模型、债券数据分析、MLflow、Airflow、Jenkins、CI/CD、OpenShift、Prometheus、Evidently AI

5G流量预测与异常检测系统　算法工程师　2021.06-2021.09

本项目旨在构建一个分布式时序预测和异常检测系统，利用 Spark 进行大规模数据处理和分布式机器学习建模，并结合 Elasticsearch 实现实时异常检测。系统从 Elasticsearch 提取海量 5G 网络数据，构建历史学习、增量学习和预测模块，提升流量预测精度，并通过动态异常检测优化运维。

- 设计并实现 基于 Spark 的分布式机器学习模型，支持海量 5G 时序数据的并行预测。
- 采用 深度学习方法或回归模型 预测未来 48 小时的 5G 网络流量，优化资源调度。
- 开发 历史学习和增量学习系统，动态调整模型，提高预测精度和适应性。
- 设计 误差学习系统，利用回归分析和异常检测方法优化预测误差。
- 构建 实时异常检测系统，结合 Elasticsearch 进行动态异常识别，提高运维效率。

技术栈

- Python、Spark（分布式计算）、Elasticsearch、HDFS、LSTM、数据挖掘、时序分析

基于机器学习的用户满意度预测系统 算法工程师 2021.05-2021.09

该项目旨在开发一个系统，从多个维度预测用户对运营商的满意度。系统针对不同影响因素构建对应的机器学习模型，并基于预测结果生成不满意用户名单，供运营团队进行精准营销和用户体验优化。

职责：

● 收集并整合来自多个数据源的数据，构建高质量的数据集。

● 进行 EDA（探索性数据分析）、数据预处理和特征工程，提取关键用户行为特征。

● 实现多种 机器学习模型，包括 随机森林、线性回归、逻辑回归、XGBoost、LightGBM、GBDT，以提升用户满意度预测的准确性。

● 基于 Spark 进行大规模数据处理和模型训练，提高计算效率。

技术栈：

Python、Spark、HDFS、Hive、随机森林、XGBoost、LightGBM、GBDT、数据挖掘、数据分析

曼谷区域学习模型 数据科学家 2020.05-2021.01

内容:

1. 搭建曼谷GIS平台, 展示地区实时数据信息
2. 对历史数据用Python清洗并传入云存储
3. 搭建机器学习模型、提取显著事件
4. 连接曼谷气象局API获取实时数据

业绩:

1. 建立曼谷首个实时数据监测、预测、自动化报警系统
2. 建立首个机器学习预测区域

奥克兰水质数据分析模型 数据科学家 2020.02-2021.01

内容:

1. 通过清理并核对所有历史数据，检验实时数据准确性，搭建自动检测系统
2. 通过azure function、Python搭建自动实时数据检测系统云函数，自动处理异常数 据，搭建可视化平台
3.建立异常数据报警系统

业绩:

1. 大幅度提高模型准确性
2. 创新性开发全套自动化检测及预警系统
3. 实现自动邮件短信报警通知需求

日本、越南等地区数据可视化模型 数据科学家 2019.10-2020.12

内容:

1. 运用云函数实现连接客户API获取、处理并导入数据，建立数据监测分析模型平台
2. 运用云函数爬虫爬取并维护气象信息数据
3. 按客户需求开发实时气象信息分析及预测模型(回归分析、最优风切模型、weibull仿真模拟等)
4. 开发可视化面板

业绩:

1. 实现客户全程自动化、全托管的需求
2. 实现客户各项分析模型的需求
3. 得到客户认可并追加能源可视化分析项目

新西兰、英澳等降雨监测、预测模型 数据科学家 2019.05-2020.12

内容:

1. 搭建GIS平台、负责开发数据预测系统，建立时序预测模型
2. 开发相关联实时数据检测系统，对监测点的数据建立时间序列模型
3. 根据数据进行水位、流量、降雨量等分析模型
4. 搭建各项数据全套预警系统

业绩:

1. 提前完成项目，准确实现客户需求
2. 建立预警模型模板，优化了检测系统模型
3. 提高模型性能67%

神经网络模型进行图像识别 数据科学家 2020.02-2020.03

内容:

运用TensorFlow, Sk-learn 搭建6层CNN, RNN模型

业绩:

建立训练准确率98.7%测试准确率80.6%的模型

教育经历

坎特伯雷大学 本科 计算机科学 2017-2018

西南财经大学 硕士 会计 2012-2014

电子科技大学成都学院 本科 电子信息工程 2006-2010