

孟晓

经验：5+

求职意向：数据科学家、NLP、算法工程师、AI Engineer



santochaoya@yahoo.com.hk



+86 17628072710



https://www.linkedin.com/in/xiao-meng-851b89153



https://github.com/XimeCraft?tab=repositories

自我介绍

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。

工作经历

EPAM System (Singapore) 数据科学家 2024.03 - 至今

EPAM System 亿磐系统科技（成都）有限公司 数据科学家 2021.09 - 2024.03

- 构建端到端债券时间序列预测模型，利用时间序列分析、深度学习和机器学习
- 开发基于新闻的 NLP 解决方案，包括 Transformer 模型和 GenAI 应用
- 运用假设检验、时间序列分析、生存分析和 A/B 测试，进行市场统计分析
- 搭建 WebSocket 与 RESTful API 实现实时数据处理
- 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

亚信科技(成都)有限公司 算法工程师 2021.05 - 2021.09

- 构建与优化机器学习模型，用于用户体验评估、客户满意度优化与风险管理
- 基于电信基站时序数据，开发定制化机器学习模型与多步预测算法，提升预测精度与异常检测能力
- 多个机器学习建模全流程开发，包括数据处理、模型训练与优化部署，涉及时序预测、异常检测、增量学习与用户满意度分析

Mott MacDonald (New Zealand) 数据科学家 2019.05 - 2021.01

- 数据挖掘与处理，集成 API、爬虫、AWS S3、GIS 平台、ETL 与云函数（Azure Function, CI/CD）
- 构建时间序列模型与机器学习模型，实现预测与分析
- 异常点处理、显著事件提取分析，结合 Power BI进行数据可视化与面板开发

四川省丝绸科学研究院 会计 2012.05 - 2015.09

成都优力微波通信有限公司 硬件设计、测试 2010.09 - 2012.03

教育背景

坎特伯雷大学(新西兰)

2017.07 - 2018.12

计算机科学(本科)

西南财经大学

2013.03 - 2014.12

会计(硕士)

电子科技大学成都学院

2006.09 - 2010.07

电子信息工程(本科)

技能

● NLP & Generative AI:

实体提取，实体匹配，文本分类，情感分析，文本处理, Transformer, BERT 和 GPT 系列模型，Generative AI, Prompt Engineering, RAG, FAISS, LangChain, LoRA, Azure Cognitive Search, GenAI 评估优化, Vertex AI

● 模型预测 & 分析:

时间序列预测, 分类预测, 时序模型 (ARIMA, LSTM, Prophet, Seasonal Decomposition), 机器学习, 深度学习, 数据挖掘, 数据分析, 统计分析, 数据可视化, 推荐算法, 异常检测, A/B 测试, 特征工程

● 机器学习工程:

Airflow, MLflow, Data Drift, CI/CD Jenkins, OpenShift, Docker, Evidently AI, ETL Spark, Azure AutoML, Azure OpenAI, AWS

本项目构建了一套自动化 **金融新闻分析与价格预测系统**，涵盖 **新闻抓取、实体提取、情感分析、主题建模、价格预测** 等多个 NLP 任务，结合 LLM、LangChain 和本地知识库，探索新闻对债券价格波动的影响。

- **文本采集清洗**: 开发自动化爬取脚本，从订阅邮件账户提取新闻文章，进行结构化解析与清洗。
- **NER**: 基于 Transformer 模型（如 BERT、RoBERTa）训练 NER 任务。
- **实体匹配**: 利用 BERT 微调 + 向量检索（FAISS），模糊匹配，将提取实体与数据库对齐。
- **RAG**: 基于 LangChain 搭建 RAG 框架，结合本地知识库。
- **主题建模与情感分析**: 采用 BERTopic + FinBERT + LLM，提取主题并评估市场情绪。
- **Prompt、LLM设计优化**: 评估、优化 LLM 处理金融文本的能力，提高生成质量与规范性
- **统计分析**: 价格走势分析，假设检验、生存分析、回归分析、A/B 测试

基于 LSTM 的债券价差预测与自动化部署系统

数据科学家，机器学习工程师

2021.12 - 2022.10

探索、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統确保长期稳定性。

- **数据处理**: 收集并清理文本、时序数据，EDA 和特征工程。
- **模型开发**: 实验 LSTM、XGBoost、LightGBM 预测债券价差，利用 MLflow 进行版本管理、超参优化与模型评估。
- **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。

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。
- **债券 Skew 计算引擎**: 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
- **高并发 API 开发**: 采用 Flask + RESTful API 支持交易数据查询、库存分析、Skew 计算等关键功能。
- **数据自动分析**: 开发 Auto-Profiling 工具，实现 EDA、特征工程和数据可视化 自动化。

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。

- **分布式时序预测**: 基于 Spark 并行计算，训练大规模 5G 网络流量预测模型，提升处理效率。
- **机器学习与深度学习建模**: 采用 深度学习（LSTM）或回归模型（XGBoost、LightGBM）预测未来 48 小时流量，优化资源调度。
- **增量学习与模型自适应**: 构建 历史学习 + 增量学习 流程，动态更新模型，提升预测精度。
- **误差学习优化**: 基于 回归分析 + 异常检测，分析预测误差，优化模型稳定性。