# Peixin XU

Hong Kong: +852-6085 9429 | 23087394g@connect.polyu.hk
China Mainland: +86-18510727955 | peixinxu1999@hotmail.com
Harbourview Horizon All-suite Hotel 12 HUNG LOK ROAD KOWLOON HUNG HOM, Hong Kong, China

## EDUCATIONAL BACKGROUND

### The Hong Kong Polytechnic University (PolyU)

***Major degree***: MSc in Information Technology, Natural Language Processing Stream 09/2023-present

- **Major courses:** Artificial Intelligence Concepts, Natural Language Processing, Data Structures and Database Systems, Human Computer Interaction, Internet Infrastructure and Protocols, etc.

### University of Science and Technology Beijing (USTB)

***Major degree***: B.M. in School of Economics and Management 09/2018-06/2022

- **GPA:** 3.56/4.0 **Major GPA:** 3.78/4.0
- **Major courses:** Data Visualization (4.0), AI Concepts (3.8), Machine Learning (3.8), Discrete Mathematics (3.8), Artificial Intelligence (3.7), Parallel and Distributed Computing (3.7), etc.

***Double degree***: B.Eng. in School of Computer & Communication Engineering 09/2018-06/2022

- **Major courses:** Data Structure, Introduction to Artificial Intelligence, Machine Learning, Operating System, Principle of Computer Composition, Wireless Network Principle.

## CODING SKILLS

- ***LargeLanguagModel*** (i.e., PT&SFT on ChatGLM/Baichuan/Qwen, Refactoring in Llama-Efficient-Tuning)
- ***MachineLearning*** (i.e., Decision Tree, Random Forest, Ensemble Learning, Nearest Neighbor Method).
- ***DeepLearning*** (i.e., Attention-based Model, Embedding-based Retrieval, Actor-Critic, Q-learning).
- ***CloudComputing*** (i.e., Hadoop, Spark, PySyft, Pygrid).

## RESEARCH OUTCOMES

- ***Inrelevant Publications***

Peixin Xu, Xiaohui Li. Online education strategy design based on learner profile of analysis[J]. New Business Weekly, 2020(14):179-180,182.

Mingting Kou, Yifan Yang, Peixin Xu, et al. Favorableness or advorsity? Quantitative Research on tax policy portfolios' impact on corporate R&D manipulation[A]. China Technology Economy Forum[C], Huhhot, 2021.

Mingting Kou, Yifan Yang, Peixin Xu. Research on the change and impact of R&D manipulation of listed companies under dual tax preference policy[A]. The 16th Annual China Science and Technology Policy and Management Academic Conference[C], The Chinese Association of Science of Science and S&T Policy Research, Beijing, 2021.

## RESEARCH EXPERIENCE

**Distributed Reinforce Learning based on PyGrid Network** 07/2022-08/2022

**Self-Motivated Project**

***Core contents:*** Implemented a PyGrid gateway for federated learning and distributed machine learning.

- **Centralized Data (Mninst Dataset) and Distributed Training**: Utilized CNN network and Paillier Encryption Algorithm to achieve 91% accuracy in 3000 epoch.

***Core contents:*** Adapted Asynchronous Advantage Actor-critic (A3C) network in tft android game.

- Designed a **CNN-Attention-based Actor Network**, and **value-based CNN Critic Network** estimated by **temporal different (TD)** methods; Server update and return the general model through local gradients.
- Designed a mouse operation actor to achieve movement, character deployment and character trading.
- Designed a state estimator system including a hit points predictor, a reward collector and an effective supervisor via **CNN** and **Hierarchical Clustering**.

**Serious Game Design for Children Autism Treatment and Evaluation** 12/2020-12/2021

**Core Member** **Supervisor: Prof. Huansheng NING**

**National University Student Innovation and Entrepreneurship Program**

***Core contents:*** Based on the theory of computer game assisted therapy and the current autism treatment, proposed an emerging serious game design for children autism therapy and evaluation.

- Adapted **Facial Recognition Model** and **Gesture Recognition Model** pretrained by *Baidu* to design an interaction game for enhancing autism kids with emotional expression and expression recognition abilities.
- Finished the **Thread** connection and data transmission with the server, using ***Python Socket*** for terminal TCP protocol connection, eventually reaching an average of **133** data records per user within 10 connections.

**Tax Preferences and R&D Manipulation: Listed High-tech Companies** 10/2020-03/2022

**Core Member** **Supervisor: Prof. Mingting KOU** **National Natural Science Foundation of China (NSFC)**

- Established request architecture crawler by **URLLIB** (Python) for listed companies from *TianYanCha*; verified and cross-validated with official disclosure data. A total of **38,200** data were obtained.
- Adopted Python ***Pandas*** and ***Matplotlib***, and ***Pyecharts*** for raw data cleaning; obtained **12,325** records from **1,098** eligible companies.
- Retrieved journal publishing data (as research outputs) of eligible companies from CNKI by ***Python Selenium***.

## INTERNSHIP EXPERIENCES

**Guangdong Oppo Mobile Telecommunications Corp., Ltd. (OPPO)** 05/2023-09/2023

**MLE Intern** **Mentor: Mr. Yuan Yuqing**

***Core contents:*** User **l**ife**t**ime **v**alue prediction model design for users' payment prediction

- Predicting User Payment in 30 days by statistic features and user portraits via Light-GBM model.
- Conducted a 2-stage model for 90-day prediction by identifying low contribute user before regression.

***Core contents:*** User **Po**sition of **I**nterest prediction via LLM (ChatGLM-6B)

- Task redefined from Multi-Classification to Retrieval-based history selection to avoid LLM Hallucination.
- Conducted the **supervised fine-tuning** with sequential POI and user portraits, with 68% Accuracy (base-33%).

***Core contents:*** Game Recommendation via LLM (Qwen-7B/Qwen-13B)

- **Pretrain** data with game definitions, similar game recommendation and the most similar game selection.
- **Supervised fine-tuning** Qwen model with target game prediction task of each user via user portraits features (beats current LightGBM model with 8% improvement in AUC).
- Ensembled with a separate prediction model for sequential data. (achieved 10% improvement in AUC together)
- Training framework refactoring for prediction possibility outputs of Binary and Multi-class classification tasks.

**Beijing Dajia Internet Information Technology Co., Ltd. (Kuaishou)** 08/2021-04/2022

**MLE Intern** **Mentor: Mr. Chi Cheng**

***Core contents:*** User lifetime prediction model design for users' retention prediction

- Adopted **routine time-series data stream** implemented by **MapReduce** to analyze features of different users during reduce stages and generate static features (i.e., sum, max, avg.) and **sequence features**.

***Core contents:*** Increase the auditing/filtering speed of local videos with proper copywriting and overseas promotion

- Utilized the filters of language, text length and regex-based content (i.e., **stop-word-lists**) and semantic similarity sorting (**Google-BERT model**) to identify videos with proper copywriting via **Spark (Scala)**.
- Filtered videos were promoted overseas by language translation of video copywriting via the **grpc interface**.

***Core contents:*** Offline data stream development

- Optimized the offline calculation code of video push for user-creator interactions through **data disassembly** to achieve **tens of billions level aggregation** for 90-day data, increasing ~3-fold calculation speed.

**Guangzhou VIRDYN Network Technology Ltd.** 08/2022-03/2023

**MLE Intern** **Mentor: Mr. Zhang Xiang**

- Conducted the Facial Reconstruction, Pose Estimation and Meta Meeting Room based on **MediaPipe**.

**Beijing Xiaoju Technology Co., Ltd. (Didi)** 06/2021-08/2021

**DA Intern** **Mentor: Miss Sun Siyi**

- Producing & Maintaining & Analysing data relevant to User-Growth & Driver-Growth though Spark & SQL.

## EXTRACURRICULAR EXPERIENCES

**The 70th anniversary of the founding of the People's Republic of China** 07/2019-10/2019

**University of Science and Technology Beijing Freshman Volunteering Activity** 07/2019-08/2019

# 徐培鑫

13719024566 workshoprd@163.com
https://xcanton.github.io

THE HONG KONG POLYTECHNIC UNIVERSITY 香港理工大學

## 教育经历

**香港理工大学** - Natural Language Processing 硕士 Department of Computing 2023.09 - 2024.06
**北京科技大学** - 大数据管理与应用 本科 经济管理学院 2018.09 - 2022.06
**北京科技大学** - 物联网工程 本科 计算机与通信工程学院 2018.09 - 2022.06

## 专业技能

- 熟练运用**Python**，能够编写**SQL**、**Spark**等程序，熟悉常用**Linux**命令和常用**Git**命令。
- 熟悉分词、情感分析等自然语言处理方法，拥有处理序列特征、时序特征的能力与经验。
- 了解Pytorch等深度学习架构，具有实现既定机器学习、深度学习模型并部署推理服务的经历。
- 具有大语言模型预训练与**Prompt tunning**微调的经历，能够针对既定模型进行探查和设计与训练任务。

## 实习经历

**OPPO广东移动通信有限公司** - 算法实习生 数据智能研究院 2023.05 - 2023.09

*主要职能：通过传统模型、深度模型、大语言模型为用户增长模块赋能，在用户价值、用户选择方面提供数据支撑。*

- **用户价值（Ltv）预估模型**
  - 通过安装、卸载、启动、付费等统计特征，基于LightGBM得到预测30天价值（ltv-30）并迭代**两阶段**ltv-90模型。
- **用户兴趣点（PoI）预测模型**
  - 基于base深度模型和数据流进行数据重构，将任务建模成有召回任务（历史+近邻）并设计prompt语料。
  - 对基模型ChatGLM探查得出：模型对地理信息具有一定的理解能力，并具备从候选集根据下标选择的能力。但经纬度计算、时间比较能力较弱。根据此设计与训练任务：时间字符串比较、固定格式输入的数据提取、经纬度距离比较等。
  - 较base深度模型（33%）：无地理信息直接训练提升**7%**，加入预训练任务还提升**15%**，加入地理信息再提升**13%**。
- **新游用户增长专项**
  - 圈选高价值用户任务中，在训练集下采样10%的情况下，ChatGLM+ptuning提升1.4%、ChatGLM+Lora提升1.9%。
  - 针对游戏领域内描述设计预训练任务：游戏理解、游戏类别预测、相关游戏推荐、相关领域百科知识注入。
  - 构建统计模型和时序模型（Qwen），通过模型集成的方式实现用户在新游拉新、老游拉新和老游拉回流任务。

**北京达佳互联信息技术有限公司（快手）** - 推荐算法实习生 KIBT-海外推荐组 2021.08 - 2022.04

*主要职能：基于深度学习模型、预训练表征模型、Spark计算架构、Hive和Redis存储架构为快手海外App提供推送消息。*

- **用户生命周期预测模型**：
  - 通过时序模型对用户在过去30天的推送干预下，预估n天内留存情况与概率。利用**Tensorflow**实现部分网络特征。
  - 基于**MapReduce**实现例行样本数据流，利用**shuffle**和**reduce**阶段处理用户状态，产出文本特征（活动标签、最长词、最常用词、最长用表情）、用户特征、序列统计特征（求和项、最大值、均值）和序列特征。
- **基于Spark的待审核推送自动化生产**：
  - ***自动化文案筛选***：为简化候选推送的撰写审核流程、避免人工挑选视频和构思相应文案，通过对近期优质视频及其用户文案，经多种规则过滤（有效性过滤、基于分词与词表的内容过滤、基于**BERT**的语义相似度排序）产出待审核推送。
  - ***自动化文案改写***：对于海外不同运营地区间有效候选推送、推送审核速度不均衡的情况，通过Spark调用GRPC模型接口，将部分语种相近区域的推送进行翻译改写，根据规则过滤后产出优质待审核推送。
- **离线数据流**：
  - 基于**Scala**更新双塔推荐模型的**Spark**数据流，新增部分特征（创作者特征、用户交互特征）。更新和新增离线数据流。
  - 通过**数据拆解**（天级、周级、月级）实现**TB**级数据90天聚合的优化，减轻计算时长约70%。

**北京小桔科技有限公司（滴滴）** - 数据挖掘实习生 IBT 2020.06 - 2021.08

*主要职能：通过Hive-sql、python实现基于Spark框架的数据生产与消费，提供数据可视化并协助分析师完成分析报告。*

- **数据生产**：通过**Hive-sql**对数仓原始log数据进行读取加工，例行产出司机与用户增长数据、渠道获客成本数据。
- **数据消费**：通过**sql**实现看板数据计算并存入Clickhouse，提供司机与用户的天级、周级和月级数据看板。

## 项目经历

**毕业设计-基于微博评论的舆情导向性推荐算法研究** - （MAE:0.001） 2021.12 - 2022.04

- 通过爬虫收集90+微博推文的10w+评论转发数据，解析传播结构并构建微博转发社交网络，并获取用户的用户信息。
- 借鉴**双塔推荐模型**结构，通过神经网络预测文本对受众的情感影响率（**BERT-embedding**余弦值vs随机负样本，1:1）。

**毕业设计-基于GRU和Self-Attention结构的循环神经网络音乐推荐算法研究** - （Top1 Acc:48%） 2022.02 - 2022.06

- 基于Million Song Dataset构建模拟音乐消费序列，通过原相似音乐家、创作歌曲数据构建消费序列。
- 通过点击率预测的神经网络预测用户对音乐的点击率，利用**GRU**单元处理*近10次音乐消费*序列特征，Top1 Acc达到48%。

**基于Pygrid的强化学习网络训练框架** 2022.07 - 2022.08

- 基于**pygrid**实现支持**pysyft**协议访问的联邦学习网关，Mnist手写数字Cnn网络预测准确率达到91%。
- 在**A3C**网络的基础上实现以Cnn网络作为共享输入网络、以注意力机制为Actor和以mlp作为critic的强化学习模型，并且在pysyft协议下改写其训练规则：本地训练计算当前批次梯度，在服务器端进行梯度汇总与模型更新，并回传训练后模型。

## 竞赛经历

**美国数学建模比赛 Interdisciplinary Contest In Modeling** - 队长 国家二等奖 Honorable Mention
**“长风杯”大数据分析与挖掘竞赛** - 队长 国家级奖项提名，华北赛区一等奖