Zheyan Liu

(929) 202-6721 | zheyan.liu@columbia.edu | New York, NY | https://zheyanliu99.github.io/

DATA SCIENCE / DATA ANALYST INTERN

Columbia University MS in Biostatistics/Data Science (2023) with hands-on experience via two Data Science/Analyst internships and research assistant positions in health-related project.

- Data science toolkit includes: Python (pandas, sklearn, pytorch), R (tidyverse, rshiny), SQL, SAS, Azure, AWS, Spark
- Substantial experience with fundamentals including statistics, feature engineering, machine learning, deep learning
- Analysis/research experience includes:
  - Leveraged GNN model to build an application that indicates the most crime-free travel routes in NYC
  - Used NLP to classify social media user sentiment during Covid; BERT model achieved 0.86 AUC.

EDUCATION

Columbia University - Mailman School of Public Health, New York, NY — Expected May 2023
Master of Science in Biostatistics, Specialization in Data Science
Relevant Coursework: Data Science (R), Biostatistics Method

Shanghai University of Finance and Economics (SUFE) - School of Statistics and Management, China — Jun 2021
Bachelor of Science Statistics
Relevant Coursework: Data Mining (Python), Machine Learning, Database Theory (SQL, C#), Programming (C++), Mathematical Analysis, Probability, Advanced Mathematics, Regression Analysis (R, SAS), Random Process, EDA (R)

PROFESSIONAL EXPERIENCE

Ping An Technology (Shenzhen) Co. Ltd, Shanghai, China — Apr 2021 – Jul 2021
Data Scientist Intern

- Implement and deploy data pipelines, data models and data processes to production with PySpark and Apache Zeppelin
- Optimized web crawlers code and feature engineering process; **improved efficiency by 40%**
- Used LightGBM, XGBoost to predict if customer will renew auto insurance and health insurance with **AUC 0.93**; Research the feasibility of implementing the GNN auto insurance anti-fraud model

Shengqu Information Technology (Shanghai) Co. Ltd, Shanghai, China — Oct 2020 – Apr 2021
Data Analyst Intern

- Used SQL to query data and Python to **automate reporting**, with Python and visualization with Plotly and Tableau; **Reduced reporting time** from hours to seconds.
- Researched on user payment/churn issues and provide in-depth data reports to the game planner teams
- Feature engineering with Kmeans and RFM; Ensembled XGBoost, Random Forest to predict user churn with **AUC 0.92**

PROJECT EXPERIENCE

Azure Data Factory project on Covid19 | Team Leader, New York, United States — Dec 2021 – Jan 2022

- Created Data Flows to ingest and transform data from sources as HTTP; Scheduled pipelines with triggers to update daily
- Leveraged Azure Monitor and Log Analytics to monitor pipelines and used Power BI to build a reporting dashboard

NYC Subway Crime GNN Prediction and Application | Team Leader, New York, United States — Nov 2021 – Dec 2021

- Cleaned and imputed 10 years of New York City subway passenger data; Built project website and organized report
- Conduct EDA on subway passenger flow and location; Clustering station coordinates into 8 clusters with K-means
- Built subway crime navigation **Rshiny App** based on Google Map Api and adapted GNN to predict crime in each route

Sentiment Analysis of Covid19 Internet Public Opinions | (Graduation Thesis), Shanghai, China — Apr 2021 – May 2021

- Used methods such as back translation, Bert synonymous replacement, and random replacement for **data strengthen**.
- Establish and tune BERT and FastText-based SVM and XGBoost models to predict sentiment tendency with **AUC 0.89**
- Applied voting model to the crawled epidemic-related Weibo data, and analyzed online public opinion of Covid

VOLUNTEER EXPERIENCE

Team Leader, **Volunteer Dept of SUFE School of Statistics and Management,** Shanghai, China — Sep 2017 – Jun 2019
Leader, **Field Research & Volunteer Teaching in Sichuan Liangshan Yi Autonomous Prefecture**, China — Jul 2019