Hanyu (Yuki) Liu

Personal Website | +1 (646)-290-1094 | hanyuliu041@gmail.com | New York, NY, 10003

WORK EXPERIENCE

MailOnline - *AdTech* New York, NY
Data Analyst (Data Engineering focused) Jun 2024 - Current

- Built and optimized ETL pipelines to pull data from APIs into database using Airflow DAGs, ensuring integrity with validation and error handling. Monitored 100+ daily DAGs to guarantee reliability and scalability in line with SLAs
- Migrated Airflow deployment from local to Kubernetes on GCP using Docker for containerization and Git for version control, improved workflow efficiency by 40% and optimizing compute resource usage by 5%
- Improved e-commerce campaign performance visibility by 30% by designing an ETL process to integrate data from multiple sources, designing data models and building Tableau reports to track key metrics like conversion rate and CTR
- Launched Home Page Block Real-Time Tracker Dashboard with alerting features, enabling timely detection of impression drops by implementing geo-specific thresholds based on hourly data patterns
- Monitored global commercial reporting for multi-million dollar business, ensuring accuracy and timely delivery

TomTom - *Maps Technology* New York, NY
Data Scientist Intern Oct 2023 - Jan 2024

- Enhanced map search user satisfaction by 2.1% and accelerated weekly data delivery by 73% and by designing and developing automated tagging of 1M geographical locations using LLM BERT transformer
- Expedited data ingestion time from 1 week to 1 hour by mapping supplier data headers to company standards using machine learning classifiers (Random Forest, XGBoost), improving model performance by fixing class imbalance
- Created a web scraping tool using Selenium to capture timely updates from Open Street Map community, and wrapped it in RESTful API with front-end interface to make it easily accessible to business team, saving 40% of their labor time
- Leveraged SQL to clean and transform 10K nested JSON data, building tables to enable analytics team to conduct map quality evaluation and competitive market analysis

WelSpot - *FinTech Startup* New York, NY
Data Scientist Intern Aug 2023 - Oct 2023

- Streamlined ELT processes with Airflow, integrating data from multiple sources into Snowflake to update personal loan records, set unique identifiers to ensure integrity constraints and maintain data accuracy
- Utilized Power BI to visualize loan performance metrics with 15K data, providing real-time insights into risk profiles
- Engineered a chatbot integrating OpenAI API, enabling domain-specific Q&A and web searches by prompt engineering
- Fostered Agile practices in close collaboration with non-technical stakeholders, aligning with key business goals

IEEE - *Technical Professional Organization* New York, NY
Data Science Advisor Apr 2023 - June 2023

- Enhanced search precision in research database by improving keyword extraction from unstructured full-text articles
- Developed ETL pipeline with PySpark to retrieve 500+ articles from AWS S3 to Redshift and process on EC2, achieving a 21% reduction in batch processing time via optimizing job partitioning and caching
- Conducted text preprocessing and fine-tuned a BERT model, resulting in 13% increase in database keywords quality
- Led project initiatives and client discussions, ensuring project alignment and stakeholder satisfaction

PROJECT

Substance Abuse Treatment Disparities - *Merck Datathon* Apr 2023 - May 2023

- Optimized a neural network to reach 85% prediction accuracy to identify key factors influencing treatment outcomes
- Generated a comprehensive report with 2 interactive **Tableau dashboards** to deliver actionable strategies

EDUCATION

M.S. in Data Science | Cornell University | GPA: 3.96 Sep 2022 - Aug 2023

- Coursework: Data Science (Python & R), Data Management (SQL), Machine Learning, NLP, Statistics

B.S. in Applied Mathematics | University of Nottingham | GPA: 3.93 Sep 2018 - Jul 2022

- Dean’s Scholarship for 2019, 2020, 2021

SKILLS

- **Programming: Python, SQL, R**
- **Cloud & Big Data:** AWS (S3, EC2, Redshift), GCP, Azure, PySpark, Snowflake, dbt
- **Tools & Reporting:** Airflow, Git, Unix, Docker, Kubernetes, Tableau, PowerBI
- **Libraries/Tools:** Pandas, NumPy, Matplotlib, Seaborn, TensorFlow, scikit-learn, NLTK, SpaCy, LangChain, RestAPI