Ying Liu

• **GitHub:** https://github.com/yingliu1206
• **Tel:** (341)-345-7248 • **Portfolio:** https://yingliu1206.github.io/portfolio/index.html
• **Email:** ying.liu0331@gmail.com • **LinkedIn:** https://www.linkedin.com/in/ying-liu-4b45a8195/

EDUCATION

Georgetown University **08/2020 - 05/2022**
Master of Science in Data Science and Analytics, GPA: 4.0/4.0 *Washington DC, US*
Shanghai International Studies University **09/2016 - 06/2020**
Bachelor of Arts in Business English, GPA: 3.82/4.0 *Shanghai, China*

- **Awards:** Province-level Merit Graduate (Top 1%), Leadership Excellence Award (Top 5%), Merit Scholarship (Top 5%)

SKILLS

- Programming Skills: Proficient in **Python**, **R**, **SQL**
- Big Data Technologies: Experience in **AWS**, **Azure**, MongoDB; Familiar with Hadoop, Spark
- Tools: **Tableau (Tableau Desktop Specialist)**, Git, Docker, Flask

EXPERIENCE

Kaiser Permanente **06/2022 - present**
Data Analyst, Financial and Operational Department *Rockville, MD*
Working Tools: **Python, SQL, SAS, Tableau, Excel, Git**

- Optimized database triggers and stored procedures for the budgeting tool and collaborated with the software development team to enhance the front-end functionality, improving the efficiency and accuracy of roster and panel size management
- Conducted data cleaning and exploratory data analysis on a composite dataset derived from multiple sources, and designed interactive Tableau dashboards to support regional executive decision-making on service distribution and capital planning
- Created effective Python and SAS programs to meet requests from regional specialty departments and streamline data analysis and reporting processes for internal clients, resulting in a 40% reduction in reporting time.
- Led a team-level transition from SAS to Python, resulting in a 50% increase in programming efficiency and a high-cost reduction.

Georgetown University: Massive Data Institute **09/2021 - 05/2022**
Data Engineer, Web-Crawling & Digital Data Collection Team *Washington, DC*
Working Tools: **Python**, **AWS**, **Git**, **Docker**, **MongoDB**

- Designed and implemented three end-to-end pipelines to automate the storage of images, files, and raw text scraped from websites into a MongoDB database on Docker
- Collaborated with a cross-functional team to design and initiate a large-scale data collection campaign to gather information from 100,000 major U.S. public schools
- Released an alpha version of Crawl4All, a web interface and application for automated web scraping, and documented the entire development process

Didi Chuxing Technology Company **04/2021-07/2021**
Data Analyst Intern, International Business Group *Beijing, China*
Working Tools: **Python**, **SQL**, **Excel**, **Tableau**

- Collected and analyzed 30 key performance indicators from 60 million overseas user trip records, identifying abnormal fluctuations and developing strategies to improve user experience
- Automated manual reporting processes by creating interactive Tableau reports using Python and SQL, reducing labor costs by 75% and improving efficiency
- Proactively coordinated with product managers across the organization to streamline data-related processes in product experiments, including troubleshooting A/B testing, correcting data inconsistencies, and developing new indicators

PROJECTS

Analyzing Factors for Advertisement Click Rate **04/2021-05/2021**
Utilized machine learning and cloud computing to analyze factors that impact the click rate of advertisements

- **Cloud Service Startup:** developed a cloud-based environment for analysis on **AWS EMR** and **PySpark**
- **Data Preprocessing:** processed 70 million data points stored in a S3 bucket, cleaned and merged the data tables using spark.sql
- **Feature Engineering:** applied Spark's **StringIndexer** and **OneHotEncoder** functions to encode the features including title, geo-targeting information, transformed the data by **VectorAssembler**, and established a pipeline to streamline this process
- **Machine Learning:** trained and fine-tuned **XGBoost** and Random Forest models, achieving an AUC score of 82%

Obesity Classification and Data Analysis via Machine Learning **08/2020-12/2020**
Conducted data analysis with ensemble methods to identify important factors contributing to obesity

- **Data Preprocessing:** preprocessed data using Pandas and NumPy, encoded categorical features, and normalized the dataset
- **Exploratory Data Analysis:** explored target variable distribution and relationship with factors such as age, height, weight, and calorie consumption through data visualizations using **Seaborn** and **Matplotlib**
- **Machine Learning**: trained and fine-tuned a **Voting Classifier** ensemble of high-performing models including **Random Forest, Gradient Boosting, and XGBoost**, achieving an accuracy of 95% (41% increase)