Shikai Liu

(613)483-9403 | sk.liu@mail.utoronto.ca
www.linkedin.com/in/Shikai-Liu
shikaikevinliu.github.io

TECHNICAL SKILLS

- Programming Languages: Python, HTML, JavaScript
- Data Engineering: SQL, Pandas, Dask, REST Api, GraphQL Api
- Engineering: Docker, AWS, Shell Commands, Git, Terraform
- Machine Learning: Scikit-Learn, TensorFlow, Catboost, NLTK

PROFESSIONAL EXPERIENCE

Statistical Analyst, Aviva Canada 10/2022 – Present

- Improved the current models such as gradient boosting regression/classification models (catboost) by feature engineering and hyper parameter tuning to improve the prediction performance.
- Designed **model validation** framework with statistical metrics: Residual, Feature Importance, R2, confusion matrix, and actuarial metrics: AvE.
- Implemented **data pipelines** with SQL, Python, Dataiku time & event triggered scheduler, and implemented them with **multiprocessing framework** to reduce running time.
- Provided interactive dashboards with HTML, JavaScript, DataPane.

Data Science Developer Intern, SickKids (Publication) 05/2021 – 12/2021

- Architected an end-to-end data pipeline with Python to process data from various formats, e.g. Tabular, JSON, text. Designed rules to perform outlier removal and data correction automatically.
- Collaborated with bioinformaticians and biostatisticians to integrate SickKids' data analytic platform (LocusFocus) with an open-source software (Pheweb).
- Developed scripts with **Shell** and **Selenium** to automate tasks and test functionalities.
- Developed **Python Packages** on top of PheWeb (**Flask, JavaScript**) and interpreted the genetic association testing (e.g., the p-value of Linear Regression & Logistic Regression) to attract external researchers to join the research network.

PROJECTS

Github Repo Analytics 02/2024 – Ongoing

- Orchestrated data pipeline with AWS Lambda functions, EventBridge, and Terraform to ingest Github logs data (e.g., repo forks history) into AWS S3 buckets (used as data lake) from GH Archive, and fetched attributes of the repos from Github GraphQL Api.
- Transformed the raw json data into graph-format data (nodes, edges), and conducted network analysis to identify the relationships between Github repos, and Github topics (e.g, whether ML repos and distributed-computing repos have close relations)

EDUCATION

Master of Science in Applied Computing 2020 – 2022
University of Toronto, Department of Computer Science, GPA 3.88/4
Courses: Neural Network and Deep Learning; Natural Language Computing; Machine Learning; Data Science Methods

Bachelor of Computing (Honours) 2016 – 2020
Queen's University, School of Computing, GPA 4.1/4.3
Courses: Data Structure, Algorithms, Software Development