YUNQIU (LULU) YAO

100 Haven Avenue Apt 19G, New York, NY 10032
yy2827@columbia.edu | (646)-683-6625 | yyao-lulu.github.io
linkedin.com/in/yunqiu-yao | github.com/yyao-lulu

EDUCATION

Columbia University New York, NY
Master of Science, Biostatistics (Theory and Methods track), GPA 3.97/4.0 *May 2019*

- Courses: Machine Learning, Deep Learning, Intro to Databases, Data Science, Probability, Inference
- Graduate Teaching Assistant (Spring 2018, Fall 2018)

Shanghai Jiao Tong University Shanghai, China
Bachelor of Science in Engineering, Food Science and Engineering *June 2017*

- Academic Progress Scholarship; Academic Excellence Scholarship

INTERNSHIPS

HVH Precision Analytics | Data Science Intern | New York, NY Jan. *2019 – Present*

- Write SQL to identify target population from claim database hosted on AWS (Redshift, S3, EC2)
- Create dashboards with Tableau/PowerBI on patients' demographics, diagnosis and transitions
- Use Python to perform feature selection based on univariate mutual information (MI), and construct classification models to predict potential patients (SVM, logistic regression, random forest, boosting)

Cepheid | Biostatistics Intern | Sunnyvale, CA *May 2018 – Aug. 2018*

- Constructed models on clinical data for statistical analyses (regularized logistic regression, ROC curves)
- Summarized and presented the work to the Clinical Affairs group and suggested techniques to improve

Edenred–Accentiv' | Data Analyst Intern| Shanghai, China *Dec. 2016 – Feb. 2017*

- Produced R, SQL and SAS scripts to monitor KPI changes and produce weekly/monthly sales report
- Applied clustering analysis and association rule mining in R/SAS for a recommendation engine
- Held discussions to troubleshoot problems and contributed to marketing strategies based on analyses

PROJECTS

Cancer Detection on Pathology Images with Neural Network *Sept. 2018 – Dec. 2018*

- Used Python to develop a CNN-based model for the detection of cancerous regions on a biopsy slide
- Created samples by sliding window across the gigapixel biopsy images, used data augmentation to increase sample size, and trained a model based on Inception V3 with Tensorflow/Keras
- Reached a prediction accuracy of 96.92% and F1 score of 91.53% in tumor detection and localization

Is Venmo Safe to Use? *Sept. 2018 – Dec. 2018*

- Created a database using Venmo transactions to assess the risk of information leakage for Venmo users
- Web scraped the public transaction records from Venmo API with Python, designed and constructed a PostgreSQL database, and populated it on Google Cloud Platform (GCP)
- Calculated risk score for each user, and built the front end on top of the Flask Python webserver to implement the web application for user access and interaction

Empirical Bayes (EB) Method for Haplotype-based GWAS *Jan. 2016 – Oct. 2016*

- Spearheaded the initiative to construct a haplotype-based linear mixed model with EB theory with R, applied to the genome of 1092 subjects, and inferred 3 genes associated with the trait of interest

PUBLICATION

- Chen, Z., **Yao, Y.**, Ma, P., Wang, Q., & Pan, Y. (2018). Haplotype-based genome-wide association study identifies loci and candidate genes for milk yield in Holsteins. PloS one, 13(2), e0192695.

SKILLS

- Technical: Python, R, SQL (MySQL, PostgreSQL), SAS, SPSS, PowerBI, Tableau, Git, Linux, Latex, Perl
- Machine/Deep Learning: regression, kNN, random forest, SVM, clustering, neural network, NLP, CNN
- Python libraries: keras, tensorflow, nltk, numpy, pandas, scikit-learn, matplotlib, scipy, flask, opencv
- R packages: tidyverse (dplyr, ggplot2, tidyr, readr, purr, stringr, forcats), shiny, plotly, flexdashboard