YUNQIU (LULU) YAO

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

EDUCATION

Columbia University New York, NY
Master of Science, Biostatistics (Theory and Methods track), GPA 3.97/4.0 *May 2019*
- Courses: Deep Learning, Intro to Databases, Machine Learning, 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)
- Perform visualization and create dashboard on demographics, diagnosis and transitions
- Create control population and uncover discriminative features based on mutual information (MI)

Cepheid | Biostatistics Intern | Sunnyvale, CA *May 2018 – Aug. 2018*
- Constructed models on clinical data for feature selection and statistical analyses (Lasso regression, PCA)
- Developed SAS Macros for raw datasets to streamline data manipulation, visualization and analyses
- 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, SAS and SQL scripts to monitor KPI changes and produce weekly/monthly sales report
- Collaborated on a recommendation engine with clustering analysis and association rule mining
- 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*
- Developed a model based on CNN to output a heatmap showing 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 on Venmo transactions to assess the risk of information leakage for Venmo users
- Web scraped the public transaction records and populated the database hosted on Google Cloud Platform
- Calculated risk score for users and implemented the web app with Flask to allow access and interaction

Study on the Readmission Rate for Diabetes *Mar. 2018 – May 2018*
- Analyzed 67,069 electronic medical records regarding the readmission status of patients with diabetes
- Constructed models with kNN, random forest, SVM to evaluate treatment efficacy and make predictions

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