ph2200@nyu.edu | 718-877-4720

Pan He

linkedin.com/in/seamus-he

Education

New York University, New York, NY May 2022

Master of Science in Biostatistics

- Cumulative GPA**:** 3.85/4.00
- Coursework: Regression, Survey Design, Longitudinal Analysis, Machine Learning, Deep Learning

Shanghai Jiao Tong University, Shanghai, China May 2020

Bachelor of Medicine in Preventive Medicine; Minor in Public Administration

Skills

- **Programming & Software:** Python (Scikit-Learn, Pandas, NumPy, SciPy, PyTorch, Matplotlib), R(Dplyr, Caret, ggplot2), SAS, STATA, SQL, Git
- **Analytics:** A/B Testing, Regression, K Nearest Neighbors, Boosted Tree, Support Vector Machine, Clustering, Neural Networks, Text Classification, Data Visualization, Image Segmentation
- **Certificate:** SAS Certified Specialist: Base Programming Using SAS 9.4

Professional Experience

Statistical Analyst, TigerMed bdm, New Jersey October 2022 – Now

- Data management and analysis of clinical trial data in Adam and SDTM format with SAS
- Using Python and R to find key identifiers for the effectiveness of the drugs on certain patient groups

Research Data Associate, NYU Grossman School of Medicine, New York June 2021 – Present

- Build variables' coding base (e.g., demographic, e-cigarette use) for over 60,000 observations using Population Assessment of Tobacco and Health study data, the nation's largest longitudinal study on tobacco use among youth and adults
- Analyze data using SAS, R, and STATA, including descriptive analysis, bivariate analysis, and multivariate analysis including adjusted logistic regression, adjusted linear regression, and two part model to find the longitudinal impact of e-cigarette marketing and product characteristics (e.g., flavors and device type) on tobacco use behaviors among U.S. youth and adults
- Analyze primary data of a pilot randomized controlled trial (RCT) that tests the preliminary effectiveness of a mHealth smoking cessation treatment intervention

Data Scientist Intern, Shanghai Heywhale Tech, Shanghai May 2021 – August 2021

- Analyzed the user data of Heywhale website and products by SQL and found key features to identify customer segmentation and help pinpoint target users in marketing campaigns
- Built pipelines with descriptive analysis, Random Forest, and Adaboost to analyze various datasets, which were used as theoretical and practical tutorials for the 2021 Summer Data Science Camp

Research and Projects

Deep Learning and Machine Learning Project, New York University January 2021 – May 2022

- Implemented Resnet18 and Vision Transformer(ViT) from scratch using PyTorch for image classification of natural scenes and achieved 88% accuracy by adding ResNet structure in ViT
- Used Centered Kernel Alignment(CKA) to analyze and compare the hidden layer representations of different depths to understand the relationship between and across neural network models

Movie Sentiment Classification, New York University

- Built count-based N-gram model and KenLM N-gram Model to conduct text classification with movie reviews from IMDb and used perplexity as a criterion to select the best model, and KenLM N-gram Model achieved 85% accuracy
- Compared model performance by applying RNN and LSTM smoothing on n-gram model separately to explore gradient vanishing/exploding problems with different gram length, sentence length