SOPHIE YAN ZHAO

@ yaz712@g.harvard.edu

h (412) 320-6772

a Cambridge, MA 02138

PROFESSIONAL SUMMARY

A data scientist candidate with creative modeling ideas, strong statistics and math background and the sufficient engineering skills to implement modeling ideas in an efficient way. Always eager to learn more.

SKILLS

- Python Programming and Analytics
- Distributed Computing including MapReduce, Spark
- Experience with AWS instance and Cluster
- Experience with openMP, openACC, MPI
- Experience with SQL, C programming
- R Analytics
- Knowledge Numerical methods such as floating point precision, optimization algorithms
- Sense of Data Science privacy and other ethics topics
- Intermediate Level Github
- Building Python packages with Sphinx doc
- Study of Human Decision making
- Computer vision

EDUCATION

Harvard University

Master of Science: Data Science
GPA: 4.0 / 4.0

Carnegie Mellon University

Bachelor of Science: Economics & Statistics, Mathematics
GPA: 4.0 / 4.0, Dean's List, High Honors

WORK HISTORY

Simplebet - Sport Betting Data Science Intern
New York , NY • 05/2019 - 08/2019

- Develop KPI (Key Performance Indicator) and PNL (Profit & Loss) modules in company-wise research framework
- Design python version of algorithms for game-state clustering and NBA in-game performance forecasting
- Implement distance metrics from paper
- Speed up existing codes by using multithreading tools, replacing loops with vectorized computation and deploying computation on AWS GPU instance

Harvard SEAS Capstone - Spotify Recommender System (Group) Project
Cambridge , MA • 09/2019 - Current
Use off-policy reinforcement learning to modify users' sequential decision making and build a recommender system with 150 million track records.

- Subsample using hash table and research cluster computing
- Reduce Dimension for track record variety using unsupervised learning and auto-encoder. The purpose is to combat extrapolation error in Q learning with no interaction with the environment
- Distribute model training with Pytorch

Harvard Business School - Research Assistant
Cambridge, MA • 09/2019 - Current

Harvard SEAS - Data Science Ethic (Group) Project
Cambridge , MA • 01/2019 - 05/2019
Predict authorship based on citation pattern in order to demonstrate the ineffectiveness of double-blinded review process. Provide quantitative evidence for the first time for supporting open peer-review process.

- Scraping data from PDF files Downloaded from ICML website
- Sentiment analysis; String processing
- Data was posted on Kaggle: https://www.kaggle.com/lynxwang30/authors-and-cited-authors-of-icml-paper

Kesci.com - Data Science Intern
Shanghai, China 06/2018 - 08/2018

- Designed a video course with IPython materials teaching Introduction to Data Science using Python; contents cover popular packages: matplotlib, tensorflow, pandas, numpy, scikit-learn etc
- Improve code snippets for K-lab environment(online coding environment)

Carnegie Mellon University - Linear Algebra Teaching Assistant
Pittsburgh, PA • 01/2018 - 05/2018

- Designed teaching session materials and lead discussion
- Regularly graded assignments, proctored exam and held office hours

Harvard Biostatistics Project - Gene Expression Analytics for Disease Detection
Cambridge, MA • 01/2019 - 05/2019
Designed classification model for cancer detection

- Selecting most relevant gene combinations among 10,000 candidates with respect to ovarian cancer
- Experimented with multiple classifier to select the most generalizable one, such as boosting tree, masomenos, KTSP in order to combat large variance from lab condition and inter-patient difference
- Discuss and solve the domain shift problem seen among patients record from different medical centers

Harvard SEAS Term Project - Lending Club (Group) Project
Cambridge, MA • 09/2018 - 12/2018

- Study lending club loan data of 2017 and improve loan default probabilities using model stacking, tree boosting models and other ensemble methods
- Formulate investment strategy based on predicted loan probabilities

Carnegie Mellon - Interactive Statistical Graphics (Group) Project
Pittsburgh , PA • 01/2017 - 05/2017

- Created a web-based Shiny application that presents various interactive statistical graphics on air pollutant levels in the U.S. using ggplot2, plotly etc
- Online Presentation: (Please avoid Chrome browser when opening): https://36315cmugroup15shinyappproject.shinyapps.io/315G15/

Pennsylvania Department of Health - Executive Office Intern
Harrisburg, PA • 06/2016 - 08/2016
Conducted statistical analysis on health data, and provided consultation to Medical Marijuana Committee and Health Equity Department

Carnegie Mellon University - Independent Programming Project in Pyt hon
Pittsburgh, PA • 09/2016 - 12/2016

- A 3-stage color ball-shooting game with drop-down menus, robots communicating with user and saving user information
- Written in object oriented programming with Pygame
- Youtube Introduction Video: https://www.youtube.com /watch?v=bGbCf3fqOqo

Carnegie Mellon Statistics - Neuroscience (Group) Project
Pittsburgh , PA • 01/2017 - 05/2017

- Engineered features from neural spike time series in the format of point process
- Classified neurons into different types based on spike patterns such as bursting, regular etc