Sze Pui (Sallie) Tsang

szepui.tsang@caa.columbia.edu | (201) 496 1101| LinkedIn: Sallie Tsang | Website: https://sallietsang.github.io

EDUCATION

Columbia University Mailman School of Public Health New York, NY

Master of Science, Biostatistics in Public Health Data Science (STEM) 09/2021 - 05/2023

City University of Hong Kong Hong Kong

Bachelor of Science, Applied Biology with minor in Media Communication 09/2016 - 05/2020

SKILLS

- **Technical**: R (Shiny App, Markdown), SQL, SAS, Python (Pandas, Numpy, NLTK), Tableau, QGIS, Geoda, Jupyter
- **Statical Analysis**: Survival Analysis, Natural Language Processing, Predictive Modelling, Machine learning, A/B Testing

EXPERIENCE

Columbia University

Data Analyst (Data Science Institute Scholar) 11/2022 - Present

- **Pipeline Development**: Managed and cleaned 26 biological & environmental health studies over 600K+ rows following FAIR principle for analysis and reproducibility via R & SQL; Stored at GitHub to promote data harmonization and utility
- **Dashboard**: Design and deploy 18 Shiny App for interactive analysis dashboard to facilitate investigator decision making
- **Data Analyses**: Establish hypothesis-free exploratory analyses to uncover unforeseen trends and exposure-diseases relationship by deep diving ad hoc questions according to team interest

Research Assistant 09/2022 – Present

- **Unsupervised Learning**: Apply K-Means clustering in healthy & diseased patients to explore signal-diseases relationship
- **Analyses**: Contribute to stress research with 450k+ 72 patients' blood pressure, heart & respiration rate during public speech Compare fluctuation & patterns by smoothed trend in 12 stages to identify biomarker signal changes under nervous emotion
- **Time Series**: Handled missing time series data using spline interpolation; Detrended data by differencing to remove capture association between biomarker time series; Developed transfer entropy function to detect information flow for time series

Technical Service Division, HKSAR Government 07/2018 - 09/2018

Data Science Intern

- **Supervised Learning**: Trained optical character recognition (OCR) with Supervised Machine Learning Algorithm SVM; Achieved 99.4% accuracy for OCR prediction in recognizing serial number and character of Euro banknotes
- **Data Preprocessing**: Developed deconvolution process to optimize image & extract 4000+ character; Stored extracted characters to well-organized training & testing datasets by stratified random sampling; Labelled character for classifier
- **Performance Analysis**: Tuned model with highest accuracy & shortest processing time with each banknote less than 30ms

PROJECTS

Natural Language Processing: Sentiment Analysis on Movie Review 10/2022 - Present

- Applied NLTK to tokenize 400K+ user reviews at Rotten Tomatoes and perform sentiment analysis utilizing Vader
- Analyzed relativity of features and built OLS & RF models to predict domestic gross based on Vader sentiment score
- Highlighted most important and frequent word with TF-IDF by Word Cloud to identify audience movie interest in Python

Machine Learning Project: Parkinson's Diseases Symptom Severity Prediction 03/2022 - 04/2022

- Conducted exploratory analysis checking collinearity between 17 predictor & response; removed highly correlated variables
- Built 8 Supervised Machine Learning Models (LASSO, Ridge, GAM, MARS, GBM, SVR, Random Forest) to predict Parkinson's Disease severity based on patients' age, sex, and vocal features
- Tuned model parameters, compared performance, and selected RF model with smallest 10-fold cross-validated RMSE

Data Science Project: Website and **Dashboard Building of the US Smoking Population** 11/2021 - 12/2021

- Cleaned 30k data and worked on team to build website generalizing overall 1.2% decreasing smoking trend via R
- Created 20+ high-quality deliverables for exploratory analysis describing smoking geographical & demographic distribution

Final Year Thesis, City University of Hong Kong 09/2019 - 05/2020

- Performed statical analyses 2-way ANOVA & T-test comparing effect of plastic type and concentration on mussels