Chang Shen

LinkedIn: https://www.linkedin.com/in/chang-shen-1ab140129/
Github: diana12333
Email : chang.shen@yale.edu
Mobile : +1-203-285-0457

Education

- **Yale University** New Haven, CT
  Masters of Science in Biostatistics *Sep. 2019 - Expected May. 2021*
  Relevant Coursework*: Linear Models, Time Series with R and Python, Natural Language Processing, Applied Data Mining and Machine Learning, Theory of Statistics, Bayesian Statistics, Statistical Method in Causal Inference*
- **Sun Yat-Sen University** Guangzhou, China
  Bachelor of Science in Statistics; GPA: 3.8 (Major 92.02/100); Rank Top 10% *Sep. 2014 - Jul.2018*
  Relevant Coursework*: Survival Analysis, Nonparametric Statistics, Measure Theory, Applied Regression Analysis, Complex Data Analysis(Case Study), Data Structure and Algorithm*

Technical Skills

- **Programming Languages**: Python, R, SAS, SQL, Matlab , C++, C, LaTex, Spark, Linux
- **Data Science Tools**: **R** shiny, ggplot2, dplyr, stringr, tidyverse **Python** Numpy, sklearn, pandas
- **AI Framework**: Tensorflow(Python & R), Pytorch

Work Experience

- **Acumen, LLC** Washington, D.C.
  Statistical Programming Intern *June. 2020 - Aug. 2020*
  - Optimized the zip+4 code to census tract mapping approach in claims data by introducing coordinate mapping methods and updating the outdated code to the latest version. Improved the calculation and visualization of COVID incidence and risk.
  - Built a simulation framework using a time series clustering algorithm to effectively estimate and adjust the claims delay in Part D medicare data with 0.6% error rate.
- **Yale University, Department of Computer Science** New Haven, CT
  Research Assistant *Feb. 2020 - Present*
  - **Text Summarization**: Fine-tuned pre-trained deep neural networks (BERT/RoBERTa/BART) using the WikiSum dataset and Pytorch to perform complex generation tasks, like generating structured summaries of scientific topics.
  - **Information Extraction**: Conducted large-scale text(Academic & Medical corpus) processing and analysis with Python. Applied several NLP tools, like the Facebook clinical trial parser and criteria2query, on Electronic Health Record data.
- **Hangzhou Dtdream Technology Co., Ltd** Beijing, China
  Data Engineer Intern *Sep. 2018 - Apr. 2019*
  - Led the algorithm group for the Daxing Urban Planning Project to track the population growth of the largest migration area in Beijing for better resource allocation.
  - Analyzed the correlation between the consumption of resources (water, electricity, etc.) and population growth; clustered the counties by developing characteristics with hierarchical clustering and k-means; implemented regression trees (CART) and GIS visualization using R and Python. Impacted policy makers' decision making for regulation.
- **Haolan Information Technology Co., Ltd** Guangzhou,China
  Artificial Intelligence Research Intern *Dec. 2017 - Mar. 2018*
  - **Algorithm Engineering**: Collaborated in a six-person team of algorithm engineers to develop an Android application for instant classification of the Traditional Chinese Herbal Medicine (including herb, pills, powder medicine) photos.
  - **InceptionV3 implementation**: Applied deep learning algorithms Inception V3 VGG19 with Tensorflow to classify over 120 different kinds of herbal medicine with 90% accuracy.
  - **Algorithm Optimization**: Conducted hyper-parameter tuning, optimization and regularization and realized affine transformation, Gaussian pyramid algorithm, etc to augment the limited data set and improve data set quality.

Projects

- **O Ye of Little Faith- Forecasting Old Faithful [project link]** Oct 2019 - Dec 2019
  - Collected Old Faithful geyser eruption data from the Geyser Observation and Study Association from 1970 to 2011. Refined data using regex expression in R and corrected wrong data. Conducted exploratory data analysis.
  - Applied Gaussian Mixture Model (GMM) to describe the relationship between old faithful eruption and waiting time until the next eruption and used R ggplot2 and shiny to visualize the data and GMM results.
- **Predicting Death Risk and Survival Times for the Thyroid Neoplasm Patients** Dec 2017 - Mar 2018
  - Applied Ordinal Logistic Regression (OLR) and Lasso-cox analysis to explain clinical data and define high-risk groups. Provided clinical guidance recommendations for thyroid cancer patients.
  - Implemented Semi-Parametric Cox-ph Model to explore the latent cause of Thyroid Neoplasm and used empirical survival function to perform statistical inference.
  - Achieved high accuracy in survival time prediction(Survival AUC 0.78 at 500th days).