Chang Shen

LinkedIn: https://www.linkedin.com/in/chang-shen-1ab140129/
Homepage: diana12333.github.io | **Github**: diana12333

Email : chang.shen@yale.edu
Mobile : +1-203-285-0457

Education

- **Yale University** — New Haven, CT
  Master of Science in Biostatistics — *Expected May. 2021*
  Relevant Coursework: *Theory of Statistics, Causal Inference, Time Series Analysis, Bayesian Statistics, Longitudinal and Multilevel Data Analysis, Applied Data Mining and Machine Learning, Survival Analysis, Linear Models, Deep Learning*
- **Sun Yat-Sen University** — Guangzhou, China
  Bachelor of Science in Statistics; GPA: **3.8** *(Major 92.02/100); Rank Top* **10%** — *Jul.2018*
  Relevant Coursework: *Applied Regression Analysis, Data Structure and Algorithms, Mathematical Statistics, Nonparametric Statistics, Complex Data Analysis(case study), Risk Management, Game Theory*

Technical Skills

- **Programming**: **Python** (Numpy, sklearn, pandas, SciPy) **R** (dplyr, stringr, tidyverse, caret), SAS(certificated), **SQL**
- **Tools/Software**: Linux, Git, C++, Flask, JavaScript, Matlab **Big Data:** Hive, Spark **Cloud Computing:** AWS (EC2, S3)
- **Data Visualization**: **R** (ggplot2, Shiny), **Python** (Plotly, Matplotlib, Seaborn), Tableau **Other Skills:** A/B testing

Experience

- **Yale University, School of Medicine** — New Haven, CT
  Graduate Research Assistant — *Feb. 2020 - Present*
  - **Clinical Natural Language Processing**
    - Designed and maintained an ETL process on AACT database for clinical trial eligibility analysis with PostgreSQL.
    - Redesigned Facebook clinical trial NLP parser and criteria2query to use on MIMIC-III clinical database.
    - Built an end-to-end named entity recognition pipeline for Electronic Health Record to support pre-screening and recruitment of clinical trial subjects with Python.
  - **Echocardiographic Clinical Data Analysis**
    - Conducted large-scale clinical data visualization and statistical analysis to assess reader reporting of low Gradient Aortic Stenosis based on Echocardiographic Parameters and Sex, paper pending submission to 2021 ASE scientific session.
- **Acumen, LLC** — Washington, D.C.
  Data and Policy Analyst Intern, FDA Safety Team — *Jun. 2020 - Aug. 2020*
  - **COVID-19 Risk Surveillance**
    - Extracted and validated 4 million+ Medicare claims data from multiple sources via SAS. Conducted geographical spatial analysis to measure COVID-19 risk across tracts, improving efficiency by 20%.
    - Introduced coordinate mapping methods to optimize the zip+4 code to tract mapping algorithm in R.
    - Scripted and automated the workflow of the outdated tract code updating with a matching rate of 92%.
  - **Impact Analysis of Medicare Part D Claims Delay**
    - Investigated the delay distribution and potential factors affecting the claims delay in Medicare Part D data, helping internal stakeholders detect problems and providing actionable insights.
    - Developed a simulation framework using time series clustering to adjust the claims delay in Part D Medicare data and achieved an error rate of .6 % in SAS.
- **Hangzhou Dtdream Technology Co., Ltd** — Beijing, China
  Data Scientist, Cloud Computing and Engineering — *Sep. 2018 - Apr. 2019*
  - Improved the allocation of water and electricity in the largest migration area in Beijing by better predicting future population growth trends with multiple machine learning algorithms (CART, Random Forest, XGBoost, LightGBM).
  - Integrated daily population growth data and resource usage data from different domains and automated feature selection (LASSO) and dimension reduction (PCA/t-SNE) of data, improving the feature engineering efficiency by 10%.
  - Developed an end-to-end population growth predictive modeling system. Built interactive dashboard in Shiny to deliver findings to the policymakers.

Projects

- **Detecting Sentiment Shifts in Coronavirus Tweets[link]** — Oct. 2020 - Dec. 2020
  - Modelled text data(tweets) with fine-tuned BERT for sequence classification, LDA and conducted word frequency analysis.
  - Built a Flask web app with BERT/LDA models and interactive data visualization embedded. Deployed with Heroku.
- **Predicting Death Risk and Survival Times for Thyroid Neoplasm Patients** — Dec 2017 - Mar 2018
  - Applied ordinal logistic regression and LASSO-cox analysis to identify high-risk groups and important features affecting Thyroid Neoplasm patients' life expectancy. Provided follow up clinical guidance recommendations for 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 accuracy Survival AUC 0.78 at 500th day.