

Ziyi (Zoey) Chen

Seattle, WA | 651-246-6434 | zychen09@uw.edu | linkedin.com/in/ziyichen09

EDUCATION

University of Washington - Seattle, School of Public Health

Seattle, WA

Biostatistics (*Data Science Pathway*), Master of Science

Anticipated Graduation: March 2023

University of Minnesota - Twin Cities, College of Liberal Arts

Minneapolis, MN

Economics, Bachelor of Science; Statistics, Minor (*GPA: 3.95/ 4.00 - High Distinction*)

Graduation: May 2021

SKILLS

Software and Programming languages: R, Python, SQL, MATLAB, SAS, Tableau, Stata, Microsoft Excel

Courses: Data Science, Machine Learning, Statistical Learning, Biostatistics, Statistical Inference & Modeling, Regression Analysis, Survival Analysis, Longitudinal Analysis, Predictive Modeling, Study Designs, Clinical Trials, Data Analysis & Visualization

WORK EXPERIENCE

Kaiser Permanente Washington Health Research Institute

Seattle, WA

Principal Investigator: Susan Shortreed, Senior Investigator & Yates Coley, Assistant Investigator

Machine Learning Research Internship

January 2022 – Present

- Estimating random forest models for predicting suicide risk using electronic health records data in retrospective cohort study
- Comparing outcome sampling approaches for accurate and computationally efficient prediction of rare events
- Learning about machine learning techniques for estimating and validating clinical prediction models

University of Washington

Seattle, WA

Principal Investigator: KC Gary Chan, Professor, Department of Biostatistics; Adjunct Professor, Department of Statistics

Research Assistant

December 2021 – Present

- Voxel-based morphometry (VBM) preprocessing the Alzheimer's disease neuroimaging data from ADNI by segmenting, smoothing, and normalizing using statistical parametric mapping (SPM) in MATLAB
- Researching on longitudinal models developed from the non-negative matrix factorization procedure to analyze the high dimensional brain imaging datasets, and identify atrophy patterns of brain grey matter in Alzheimer's disease using Python

University of Minnesota – Twin Cities

Minneapolis, MN

Principal Investigator: Ju Sun, Assistant Professor, Department of Computer Science & Engineering

Research Assistant

June 2021 – August 2021

- Applied the Natural Language Processing (NLP) word embedding algorithm to encode categorical features with large cardinalities, reducing the sparsity of datasets by 24% at the same time
- Compared Random Forest and Gradient Boosting with deep neural networks on the cleaned datasets, demonstrated the average testing accuracy of deep neural networks is just 2% lower than the accuracy of classical methods on smaller datasets
- Preprocessed and standardized 120 real-world small datasets from UCI using Numpy and Pandas packages in Python

JOVE-LIFE Investment Management Co., Ltd

Beijing, China

Industry Research Analyst Intern

July 2020 – September 2020

- Construct data analysis on financial statements, competitors, and industry landscapes in R to aid in market decision making
- Built data visualizations demonstrating market share of competitors, order quantity, market size, and revenue structure
- Performed in-depth research and presented market research reports about the supply chain logistics industry

CEEC & Ping'an Equity Investment Fund Management Co., Ltd

Beijing, China

Data Analyst Intern

May 2020 – July 2020

- Reorganized comprehensive financial database of two wind energy projects and discarded 25% invalid data in SQL
- Built operational dashboards about financial data analysis, cash flow measurements, integrated cost analysis in Tableau

PROJECTS

Home Value Prediction project

December 2021

- Tuned hyperparameters of Random Forest, XGBoost, Ridge, and Lasso models using GridSearchCV respectively in Python
- Adjusted prediction model estimators to achieve a reduction in the mean absolute percentage error from 0.32 to 0.14
- Conducted data analysis and visualizations to find the top10 strongest correlation features and relationships with Sale price

Weather Prediction Project

August 2021

- Fitted linear regression models to predict precipitation, added regularization of Ridge regression and polynomial components to enhance the model performance with R^2 increase of 9% in R
- Preprocessed datasets, dealt with missing values, constructed visualizations using ggplot to evaluate weather-related variables