YEH CHAN YOO

Seattle, WA | yehchanyoo@gmail.com (personal) | yehchany@uw.edu (grad school) | linkedin.com/in/yehchan-yoo | yehchanyoo.github.io

SUMMARY

Master's student in Statistics at the University of Washington with experience in machine learning, NLP, and applied predictive modeling. Skilled in reproducible ML research and statistical analysis across domains like recommendation systems, survey data, and social analytics. Passionate about leveraging emerging technologies to build scalable, real-world solutions at the intersection of AI and user behavior.

SKILLS

- ***Relevant Languages:*** Python, R, SQL, C/C++, Git, Shell, HTML, CSS, JavaScript
- ***Programs/Platforms:*** Jupyter Notebook/Lab, RStudio, Microsoft Azure, Microsoft Excel, Google Cloud Platform, GitHub
- ***Packages:***
  - **Data Collection and Manipulation & Statistical Analysis:**
    - *Python:* NumPy, Pandas, SciPy, Xarray, Statsmodels, ArviZ, Requests, Beautiful Soup, Selenium
    - *R:* dplyr, tidyr, readr, stringr, data.table, survey
  - **Data Visualization:** Matplotlib (Python), Seaborn (Python), Plotly (Python), ggplot2 (R), igraph (R), shiny (R)
  - **Machine Learning:** SKLearn (Python), PyMC3 (Python), Pytorch (Python), Transformers (Python), XGBoost (Python), caret (R), e1071 (R), kernlab (R), mclust (R), nnet (R), Mice (R)

EDUCATION

University of Washington - Seattle *September 2024 - Present (Expected Graduation in March 2026)*

Master of Science in Statistics - Advanced Methods and Data Analysis

- ***Past Coursework:*** Advanced Machine Learning (CSE 599S), Applied Regression (STAT 504), Design and Analysis of Experiments (STAT 502), Machine Learning (CSE 546), Sample Survey Techniques (STAT 529), Statistical Computing (STAT 534), Statistical Inference (STAT 512 & STAT 513), Statistics Seminars (STAT 600)

University of California - Berkeley *August 2017 - December 2019, August 2022 - December 2023*

Bachelor of Arts in Statistics and Political Economy with a Minor in Data Science

- *Note*: Paused undergraduate education from December 2019 to August 2022 due to mandatory military service in South Korea
- ***Relevant Coursework****:* Causal Inference; Concepts in Computing with Data; Concepts of Probability; Concepts of Statistics; Data, Inference, and Decisions; Introduction to Artificial Intelligence (AI); Principles and Techniques of Data Science; Linear Modelling: Theory and Applications; Modern Statistical Prediction and Machine Learning

RELEVANT PROFESSIONAL EXPERIENCE

Mindful Conversion • Data Scientist *January 2024 - August 2024*

- Served as the primary data scientist for the marketing data analytics agency and startup Mindful Conversion, supporting and leading several of the agency's complex marketing projects in a cross-functional team with both software engineers and professional marketers
- Developed the data pipeline for the company's predictive search engine optimization (SEO) product Kixely to coordinate data mining and machine learning tasks using Python, Azure, and Google Cloud Platform -- collecting and analyzing large-scale SEO web data on ~100,000 keywords and URLs
- Led exploratory data analysis on 10 GB+ of marketing data using Python and SQL, utilizing strong technical skills and critical thinking for solving problems with key data collection and content strategy for the agency and its clients
- Authored and presented over five technical reports with visualizations that provided business insights on the marketing strategies of the agency and its clients

UC Berkeley School of Education • Student Assistant/Undergraduate Research Apprentice *January 2023 - December 2023*

- Researched and engineered large language models (LLMs) for automatic short answer grading of 200+ student answers to mathematical and statistical questions to improve grading efficiency for public school teachers with natural language processing (NLP) techniques while ensuring reliability in grading; achieved 75% test accuracy with a distilled fine-tuned RoBERTa model using PyTorch and CUDA
- Graded and analyzed written answers to statistics assessment questions from more than 500 California high school and middle school students using advanced Excel and R functions and psychometric techniques, contributing to refinements in assessment methodology

RELEVANT PROJECTS

Replication of Rec-R1 *June 2025*

- Reproduced and critiqued results from a preprint publication on Rec-R1, a reinforcement learning framework integrating LLMs with recommendation systems for sequential recommendation and product search, for the final team project in the University of Washington's Advanced Machine Learning course

Link to original paper on Rec-R1: https://arxiv.org/pdf/2503.24289
Link to our replication paper: https://yehchanyoo.github.io/attachments/2025/CSE_493S_599S_Final_Project.pdf
Link to our project Github repository: https://github.com/yehchanyoo/Rec-R1_magic

Evaluation of Imputation Methods Under Different Missing Data Conditions *June 2025*

- Simulated and compared five imputation strategies (mean, random, regression, kNN, no imputation) under MCAR, MAR, and MNAR settings using the American Community Survey's public-use microdata for New York state for the final project in the University of Washington's Sample Survey Techniques course
- Assessed accuracy across metrics (means, quantiles) to identify method robustness across data missingness mechanisms
- Presented findings at the CSSS Poster Session, showcasing applied modeling and visualization skills in handling real-world incomplete data

Link to paper: https://yehchanyoo.github.io/attachments/2025/stat529_final_project_yehchan_yoo.pdf
Link to poster: https://yehchanyoo.github.io/attachments/2025/stat529_poster_final.pdf

Inference and Prediction on Crude Diabetes Prevalence in U.S. States Based on Vegetable Consumption *May 2023*

- Analyzed CDC and U.S. Census data (200+ points) using Python to model and report causal and predictive links between vegetable consumption and diabetes prevalence across U.S. states for the final team project in UC Berkeley's Data, Inference, and Decisions course
- Discovered that increased vegetable consumption may have caused a decrease in diabetes prevalence in U.S. states with 95% confidence
- Demonstrated that random forest outperformed generalized linear regression models (Bayesian and frequentist) in forecasting diabetes prevalence in a U.S. state based on vegetable consumption, reducing train and test RSMEs by ~60% and ~25%, respectively

Link: https://yehchanyoo.github.io/attachments/2023/DATA_C102_Project_Final_Project_Submission__CD_TE_CM_YY___230512_.pdf