

Summer Xia

EDUCATION

Northwestern University

M.S. Data Science

Chicago, IL

Expected graduation: Aug.2024

University of California, Santa Barbara

Major: Statistics and Data Science B.S.

Santa Barbara, CA

Sep.2019-June.2023

Dean's list: three quarters | Honors student

SUMMARY OF SKILLS

- Programming Languages: R (packages: shiny, keras, tensorflow, etc.), Python (packages: pandas, matplotlib, seaborn, Pytorch, NumPy, SciPy, Keras, TensorFlow), C++, SQL, Javascript, Linux, HTML
- Technical Tools:
 - Databases: Alicloud, MongoDB, MySQL workbench, PostgreSQL, Dbeaver, Neo4j
 - Data visualization: Tableau
 - Other: Excel, Word, KNIME, Godot, SAS

PROFESSIONAL EXPERIENCE

Realix AI (Gen AI Leadership Coach Startup)

Data Science Intern

March.2024-Present

- Prompt engineering for our product: Fine tune training model (LLM), increased the BLEU score by 20%
- Text processing and data cleaning in Python: parsing and formatting

Fintelics

Data Analyst Intern

Oct.2022-Dec.2022

- Utilized Mark to Market model with pandas and numpy packaged in Python, developed a model for interest rate swap.
- Conducted gap analysis to propose process improvements and converted business requirements into user stories, use cases and test cases.
- Used NLP sentiment analysis to analyze the relationship between news headlines and stock prices, implemented the end product on MongoDB.

ARK.IO

Data Analyst Intern

Jul.2022-Sep.2022

- Monitored and managed AlibabaCloud database (memory and CPU utilizations, indexes, etc), retrieved and updated data using SQL queries.
- Generated timely reports on user and post data with the data table retrieved from cloud database, using R, analyzed the user demographics and visualized users locations with ggmap.
- Determined the most active users and created banners for them to boost user interaction and stored the data to cloud for future analysis.

PROJECTS AND RESEARCH

Capstone with CalCOFI

Jan 2023- May 2023

An eDNA window into larval fish habitat, ecosystem structure, and function using CalCOFI data

- In charge of the preliminary analysis, data cleaning, and model development for 18s sequence eDNA datasets.
- Improved the overall interpretability of the data, pinpointed the issues within the data processing step and supervised the correction.
- Employed PCA to summarize the overall datasets, then used general linear model, and decision tree to build a predictive model for anchovy presence based on the eDNA.

AI Art Detector Project

Sep. 2022 - April.2023

Developed a neural network model to discern if the image is generated by AI;

- Proposed the project idea and led a group of two to four throughout the year.
- Used MaxVit to develop two separate neural networks: one for distinguishing between human and AI art, one for distinguishing the possible source of the AI art.

Survival Analysis Project

Fall 2022

Probability of world's political parties' leaders to stay in office for a certain time

Mentored by Professor Andrew Carter

- Based on previous research by Horiuchi and Liang, this project tries to find the relationship between the time of election and the length of the political leader's term.
- Used R packages survival and survminer to plot the Kaplan Meier estimate probability, performed step AIC to select variables and created a Cox Proportional Hazard model, and finally explored the recurrent event model for the data.

Time Series Project

Winter 2022

Ground level ozone in Los Angeles from 2000 - 2020

- Visualized the time series with R, identified the time series as a seasonal ARMA model.
- performed spectral analysis on the model, and forecasted the ground level ozone up to March 2022 with fair accuracy.

German Credit Prediction March 2024

Predictive models for the Kaggle German Credit Risk data.