

Bobby Zhu

📍 San Diego ✉ bobbyzhu.work@gmail.com 📞 858-291-3844 in Bobby-Zhu

Education

University of California San Diego

Sept 2021 – March 2025

Major: Data Science | Minor: Business Economics

GPA: 3.88/4.0

- **Coursework:** Data Management, Systems for Scalable Analytics, Machine Learning, Deep Learning, Representation Learning, Recommender Systems, Data Visualization

Experience

California Policy Lab

San Diego, CA

Data Analytics Intern

June 2024 – August 2024

- Developed an ETL pipeline for **800GB of consumer credit panel** data using **MySQL** and **R**, reducing data processing time by 30%.
- Built **linear regression** and **difference-in-difference** models to evaluate the long-term impact of natural disaster compensation on victim credit performance, supporting data-driven policy decisions.
- Conducted a literature review and adapted existing methodologies to analyze wildfire victim compensation data for a novel use case.

Scripps Research Translational Institute

San Diego, CA

Data Engineer Intern

June 2023 – June 2024

- Engineered a **data processing pipeline** for 80GB cardiac image, including conversion, segmentation, and feature extraction using a **convolutional neural network**, achieving **91% pixel-wise segmentation accuracy** and providing key features for a predictive model in cardiac disease diagnosis.
- Trained a predictive model using **XGBoost classifier**, implementing cross validation and hyperparameter tuning to predict cardiac disease from imaging data, achieving 82% accuracy and 77% AUC.
- Conducted **15 unit tests** to validate data processing and transformation methods, integrating the functions into the **Torkamani Lab repository**, streamlining workflows and significantly reducing data processing time for colleagues.

IMD Business Institute

Remote

Data Analytics Intern

July 2023 – Jan 2024

- Implemented a business prediction index by applying sentiment analysis to corporate news data from six industries using **Sklearn**, **Pandas**, and a fine-tuned **BERT model**, enabling accurate forecasts of industry trends and improving strategic decision-making.
- Automated multi-source data aggregation by integrating **Python** scripts with **4 external APIs**, optimizing the collection of news and stock data, and reducing processing time by 70%.
- Designed interactive dashboards to visualize **key performance indicators (KPIs)** and predictive trends for stakeholders using **Plotly**, enhancing report delivery efficiency by 40%.

Research

Data-Driven Suicide Prevention Initiative (PRISM)

San Diego, CA

with Dr. Danks & Dr. Wayne

Sep 2023 – Present

- Designed a modularized data integration pipeline to automate ETL processes for 500+ files from diverse sources, streamlining data preprocessing and transformation.
- Led **explainable AI efforts** using **Random Forest** and **SHAP values** to identify key features from **UMAP embeddings** of suicide data, enabling actionable insights from reduced-dimensionality data.

Teaching Experience

Halcioğlu Data Science Institute

San Diego, CA

Undergraduate Teaching Assistant

Jan 2023 – Present

- Assisted **800+** students by hosting office hours and discussion sections, explaining concepts in **Python**, **data structures**, **algorithms**, and **PostgreSQL** through visual and verbal communication.
- Automated testing for **70% of assignments** using **Pytest**, developing over **20 tests** and reducing grading time by **85%**.
- Designed and presented an automated data transformation pipeline leveraging **AWS S3**, **Lambda**, and **RDS**, streamlining data processing workflows for instructional purposes.

Projects

AWS Cloud Data Warehouse for KPI Analysis | MySQL, Redshift

Aug 2024 – Nov 2024

- Developed an **ETL workflow** to extract data from **MySQL database** and **S3 bucket**, transform it with **AWS Glue**, and load it into **Redshift**, reducing data processing time by **40%**.
- Created **7 SQL scripts** to automate KPI calculations, including daily sales trends and monthly revenue growth, improving forecasting accuracy by 20% and enabling real-time decision-making.
- Leveraged **Apache Superset** to visualize KPI results, creating interactive dashboards that provide actionable insights for stakeholders.

Latent Class Classification for Sepsis Patients | Python, PostgreSQL, R

Sep 2024 – Present

- Configured a **Dev Container** for a team of six, integrating Miniconda and PostgreSQL containers to establish a consistent and efficient development environment.
- Developed 10 SQL scripts to preprocess patient groups and calculate critical illness scores, including SOFA and LCA ICU metrics, as part of preliminary cohort analysis.
- Leveraged the **poLCA** library to perform Latent Class Analysis, validating results with **logistic regression models** and achieving AUC scores above 95% for all classes.

Triple C Club Website Development | React Native, Git

Jan 2023 - Apr 2023

- Developed **4 reusable React Native components** for the Triple C Club's static website, enhancing modularity and maintainability of the frontend.
- Collaborated with a team to manage and develop features across multiple branches using **Git**, performing code reviews to ensure high-quality and maintainable code.

Extracurricular Activities

Chinese Engineering Society - Career Development Officer

Nov 2022 – Sep 2023

- Organized **7+** professional development workshops, networking events, and mentorship sessions, enhancing skills in career planning and personal branding.
- Coordinated **3 seminars and 1 panel discussion** with peers and industry professionals, fostering collaboration and insight-sharing among **150+ attendees**.

Skills

Languages: Python, SQL, Java, JavaScript, R, HTML, CSS

Technologies: Python DS stack (e.g., Numpy, Pandas, Sklearn), PyTorch, Regex, PostgreSQL, MySQL, Spark, Hive, Hadoop, AWS, HPC, Linux/Unix Commands, GitHub, Microsoft Office