Stephanie Wang

(510)-918-8739 | ssyuw0418@gmail.com | ssyuw.github.io | U.S. Green Card Holder

Education

University of California, San Diego Sept. 2022 – (expect) Mar. 2026
Bachelor of Science in Data Science GPA: 3.8/4.0

- Coursework: Probabilistic Modeling & Machine Learning, Representation Learning, Data Structure & Algorithm, Search Algorithms, Prob & Stat, Data Visualization, Mathematical Reasoning, Linear Algebra, Vector Calculus

Technical Skills

Languages: Python (pandas, numpy, scikit-learn, xgboost, lightgbm, catboost, seaborn, matplotlib), SQL (Postgres), Java, JavaScript, R, HTML/CSS, JSON
Technical: Machine Learning (Tree Models, Linear Models, Ensemble Methods, Deep Learning), Data Processing (EDA, Feature Engineering, Cleaning, ETL), Statistical Analysis (A/B Testing, Hypothesis Testing, Modeling), Data Visualization (Tableau, Power BI, D3), Web Scraping
Tools: AWS, Git, Excel, Microsoft office, Jupyter lab, Matlab

Experience

Undergraduate Research Assistant Feb. 2025 – Persent
Halıcıoğlu Data Science Institute, UC San Diego San Diego, CA

- Assisted in data collection, preprocessing, and exploratory analysis using **Python** and **SQL**, ensuring data integrity while conducting statistical analysis (**hypothesis testing, A/B Testing**) to uncover key patterns.
- Supported **machine learning model** development by engineering features and evaluating model performance, optimizing hyperparameters and testing different model architectures to enhance predictive accuracy.

Data Scientist Intern Jul. 2024 – Sept. 2024
Chuling Information Technology Co., Ltd.

- Cleaned, integrated, and analyzed sales data using **SQL** and **Python** to improve data consistency and ensure high-quality insights for decision-making. Conducted data validation and anomaly detection to enhance reliability.
- Developed **Gradient Boosted Decision Trees** and **time series models** to forecast customer behavior, applying feature engineering, cross-validation, and hyperparameter tuning to optimize accuracy and interpretability.
- Built optimized **SQL queries** and interactive dashboards (**Tableau**) to visualize sales trends and customer segments, enabling automated reporting and improving data-driven sales strategies.

Projects

Rating Prediction for Amazon Products | *Machine Learning, Natural Language Processing*

- Developed an **ETL** workflow to process 490k+ text records, applying imputation, text normalization, regex-based preprocessing, and outlier detection to enhance data quality and consistency.
- Engineered **NLP** features (TF-IDF, Word2Vec, Doc2Vec) for sentiment and rating prediction, enhancing model interpretability with feature selection and dimensionality reduction.
- Built and optimized a **logistic regression model** with L1/L2 regularization, evaluated performance using RMSE, R-squared, and conducted error analysis and feature importance visualization to refine predictions.

Power Outage Risk Analysis Modeling | *Machine Learning, EDA, Feature Engineering*

- Conducted **exploratory data analysis** on geographic and climate datasets using **Python** to identify patterns and correlations in power outages. Applied **statistical analysis** to derive actionable insights for model development.
- Engineered time series and spatial features using One-Hot Encoding, Scaling, Lag Features, and Aggregation Methods to capture temporal dependencies and regional risk factors.
- Developed and optimized a **Gradient Boosting Model** for predicting outage causes, utilizing hyperparameter tuning (GridSearchCV, Bayesian Optimization) and model evaluation metrics to improve reliability and accuracy.

LLM Based Web Agent | *Large Language Model, Prompting Engineering, Computer Vision*

- Developed automated web interaction and data extraction workflows using Selenium, handling dynamic page elements, managing delays with explicit waits, and optimizing task execution for large-scale data collection.
- Applied **OpenCV** for image processing, implementing background subtraction, and boundary detection to improve target detection accuracy and enhance feature extraction in noisy environments.
- Integrated **large language models**, designing and fine-tuning prompts to automate complex decision-making tasks, while optimizing inference speed and accuracy for real-time applications.