Baixi (Steven) Guo

415-359-4897 | bxsguo@gmail.com | github.com/StevenG777 | linkedin.com/StevenGuo777

EDUCATION

University of California, San Diego **January 2025 – Expected Graduation June 2026**
Master of Data Science **Cumulative GPA: 4.00**

University of California, Merced **August 2019 - December 2023**
Bachelor of Science, Double Majors in Computer Science and Applied Mathematics **Cumulative GPA: 3.85**

SKILLS

Languages: Python, R, Matlab, SQL, Javascript, C++; **Databases:** MySQL, PostgreSQL, Prometheus, Firestore
Technical Tools: Scikit-Learn, Keras, PyTorch, NumPy, Pandas, Matplotlib, NLTK, Google Cloud, Excel

INTERNSHIP EXPERIENCE

Conectado, Inc. **Remote**
Backend Development Intern **January 2024 - May 2024**

- Spearheaded the Python web scraper automation, reducing script development time by 50%
- Implemented data deduplication and standardization, ensuring data integrity and reducing redundancy
- Initiated data migration from MySQL to Firestore, and deployed the scraper to the GCP serverless function

Open Avenues Foundation **Remote**
Machine Learning Student Consultant **September 2023 - October 2023**

- Conducted radiology medical report classification using Python, enhancing efficiency for radiologists
- Applied a logistic regression model, generating Word2Vec embedding, and achieved an accuracy of 99.7%
- Implemented T-SNE clustering using Scikit-Learn, enhancing insights for reports through cluster analysis

Open Avenues Foundation **Remote**
Data Analysis Student Consultant **July 2023 - August 2023**

- Performed healthcare data visualization using R and utilized its insight to identify relevant features
- Developed logistic regression models to predict adverse event risk, achieving an accuracy of 86%
- Improved the model's accuracy by 1.5% through feature selection using stepwise selection methods

Lawrence Livermore National Laboratory, University of California, Merced **Merced, CA**
Data Science Challenge Intern **May 2022 - June 2022**

- Implemented machine learning to screen chemical candidates for COVID drug discovery in Python
- Conducted classification employing supervised learning models from Keras and Scikit-Learn, and achieved an accuracy of 82% with Multi-layer Perceptron, demonstrating superior predictive performance
- Performed hyperparameter tuning, leading to 2% optimization in running time and prediction accuracy

PROJECT EXPERIENCE

Scalable Product Review Prediction | *Python, PySpark, Spark SQL, Spark MLlib, Word2Vec, PCA, Decision Trees*

- Leveraged a Spark cluster to process 25GB+ of Amazon product review data in a distributed environment
- Engineered 10+ features (aggregation, imputation, flattening, and encoding) across 8 ETL and ML tasks
- Built regression models to predict product review scores, enabling data-driven product quality insights

Ant vs. Bee Image Classification | *Python, PyTorch, Scikit-learn, Computer Vision, XGBoost, Image Processing*

- Automated a pipeline to normalize and resize over 450 raw images using Torch Vision transform
- Leveraged transfer learning with a pretrained ResNet50 model to extract image embeddings using PyTorch
- Utilized models such as Logistic Regression and XGBoost, with XGBoost reaching a test accuracy of 82%
- Fine-tuned XGBoost using stratified cross-validation and early stopping, boosting test accuracy by 1%

Obesity Level Prediction | *Python, NumPy, Pandas, Seaborn, Scikit-learn, PCA, Logistic Regression, EDA*

- Analyzed the UCI ML dataset to predict obesity levels based on demographics, dietary, and lifestyle habits
- Conducted extensive exploratory data analysis via descriptive statistics and visualizations
- Applied data preprocessing techniques, including one-hot encoding, ordinal encoding, and feature scaling
- Trained a logistic regression model, achieving 88% (±1.5%) average accuracy over 500 randomized trials
- Identified key predictors via L1/L2 norm analysis of model coefficients and heatmap interpretation