ANDY WANG

awang124@jh.edu • (443) 465-1358 • linkedin.com/in/awang124 • awang124.github.io

EDUCATION

Johns Hopkins University   Baltimore, MD
Bachelor of Science in Applied Mathematics and Statistics   Expected May 2027
Second Major in Computer Science   Cumulative GPA: 3.86 / 4.0

Current Coursework: Deep Learning, Computer Vision, Mathematical Statistics, Stochastic Processes
Previous Coursework: Intro Data Science, Probability, Optimization, Linear Algebra, Multivariate Calculus

SKILLS

Programming: Python, R, SQL, C/C++, Java
Data Science: Python (Scikit-Learn, Keras, Pandas, NumPy, Matplotlib), R (tidyverse, tidymodels), PostgreSQL
Miscellaneous: Shiny, Git, Bash, Jupyter Notebooks

RESEARCH EXPERIENCE

Laboratory of Computational Intensive Care Medicine   September 2024 – Present
Johns Hopkins School of Medicine | Baltimore, MD

- Queried 398 million entries of the MIMIC-III Clinical and Waveform PostgreSQL databases, identifying 17014 patients with Traumatic Brain Injury and extracting their ECG data and medical histories.
- Utilized Python and BioSPPy to correct ectopic heartbeats and interpolate low-frequency ECG signals.
- Applied spectral analysis and collaborated with my team to extract 31 heart-rate variability features, including time-domain, frequency-domain, and nonlinear measurements.
- Designed and implemented a self-supervised transformer and custom attention layer using TensorFlow and Keras to generate HRV time-series feature embeddings for clustering.
- Discovered 3 distinct TBI patient phenotypes through hierarchical clustering for use in personalizing medical treatments and achieving faster recoveries.

PROJECTS

Cardiac Arrhythmia Diagnosis   August 2024

- Applied Python and WFDB to read 24 hours of ECG signals from the MIT-BIH Arrhythmia database.
- Utilized BioSPPy to split signals into 116942 individual heartbeats and NumPy to engineer 360 features from normalized signal values.
- Performed PCA on dataset, reducing training time by 77% while preserving 95% variance.
- Ensembled and fine-tuned Gradient Boost, Random Forest, and Support Vector classifiers to create a final model yielding 0.91 F1-score in predicting heartbeat abnormality.
- Employed K-Means clustering to determine cardiac arrhythmia based on model probability output.

Apple Stock Price Prediction   July 2024

- Performed time series analysis on 2022-23 Apple stock prices using Pandas and Matplotlib, examining stationarity and autocorrelation and applying trend decomposition to identify patterns and noise.
- Trained LSTM, ARIMA, and Prophet models to forecast 2024 prices, yielding a lowest RMSE of 2.592 dollars per share compared to ground truth.
- Deployed a Shiny web application in R to forecast and visualize any company's stock using the LSTM.

Apache Spam Email Detection   July 2024

- Utilized Python, HTML, and regular expressions to parse and clean 6047 emails from the Apache SpamAssassin database, lemmatizing words and removing hyperlinks, escape characters, etc.
- Implemented word-count vectorization algorithm from scratch and built custom Scikit-Learn transformer pipeline to process emails into a sparse feature matrix.
- Trained and fine-tuned Random Forest, Gradient Boost, and Logistic Regression classifiers, yielding 96% recall and 94% accuracy in flagging spam.

California House Price Prediction   June 2024

- Conducted exploratory data analysis on 20640 California census districts with Pandas and Matplotlib, considering feature distributions and correlation to identify best predictors of median house price.
- Engineered geospatial features via K-Means clustering and the RBF kernel to reduce RMSE by 30%.
- Fitted and fine-tuned Random Forest, Ridge, and Lasso regressors, yielding a lowest RMSE of $43K.
- Compared each feature's GINI impurity decrease to determine with statistical significance that median income is the best predictor of median house price.

Northwind Sales Analysis   June 2024

- Analyzed 31700 entries from 6 tables of the Northwind Sales PostgreSQL database via R library DBI.
- Wrote complex SQL queries involving window functions, CTEs, and joins to calculate cumulative revenue, compare employee performance, identify high-value customers and best-selling products, etc.
- Created stored procedures and views to calculate total costs of orders, return products whose sales exceed a given percentile, query information schemata, etc., in order to answer difficult questions.
- Elegantly visualized engineered data using ggplot2 to communicate insights to a business audience.

OTHER EXPERIENCE

Undergraduate Teaching Assistant | Johns Hopkins University   Aug 2024 – Present

- Held weekly office hours and project workshops for over 430 Gateway Python students, fielding questions, reinforcing programming techniques, and ensuring students are equipped to succeed.
- Graded projects and homework assignments for 68 students, giving individually tailored feedback.

Positioning System for Sports Analytics | John Edison Lab   Feb – Apr 2024

- Utilized C++ and Arduino libraries to transmit/receive timestamps between DWM1000 UWB transceivers in order to measure distances between these devices with accuracy ±10cm.
- Designed and implemented a triangulation algorithm to determine the coordinates of a moving device relative to four fixed devices using distance measurements.
- Wrote Python scripts to receive positional data and timestamps from the fixed devices via serial communication, log data in CSV format, and visualize of positions over time.
- Collaborated with JHU Makerspace to build cases for the devices wearable by athletes to generate insights on various sports strategies based on players' positional data.

SAT Math Tutor | Capital Educators   Jan 2024 – Present

- Tutored 68 high school students weekly in advanced SAT math and test-taking strategies, yielding an average increase of 100 points in students' SAT math scores.
- Collaborated with supervisors and fellow tutors in planning and composing lessons and diagnostic tests.