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: 3.93 / 4.0

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 — June 2024 –
Johns Hopkins Medical Institute | Baltimore, MD

- Queried 398 million entries across 26 tables of the MIMIC-III Clinical and Waveform databases using PostgreSQL and psycopg2.
- Identified 17014 patients diagnosed with Traumatic Brain Injury; extracted ECG data, comorbidities, active medications, and treatment histories through stored procedures, window functions, and CTEs.
- Utilized Python and BioSPPy to correct ectopic heartbeats, filter arrhythmic events, and interpolate low-frequency ECG signals.
- Employed Bash scripting to extract 15 heart-rate variability features, including time-domain, frequency-domain, and nonlinear measurements.
- Discovered 7 new TBI subphenotypes through hierarchical clustering.

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 data using Pandas and Matplotlib, applying stationarity, auto-correlation and trend decomposition to determine patterns and noise within prices.
- Trained LSTM, Prophet, and Auto-ARIMA models to forecast 2024 prices, yielding a lowest RMSE of 2.592 dollars per share compared to ground truth.
- Deployed a Shiny web application to forecast and visualize stock prices from any company, any time period using the best model.

Apache Spam Email Detection July 2024

- Utilized Python, HTML, and regular expressions to parse and clean 6047 emails from the Apache SpamAssassin database.
- 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.

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

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.

Undergraduate Teaching Assistant | Gateway Computing Python Aug 2024 –

- Held twice-weekly office hours for over 240 Gateway Python students, fielding questions, reinforcing programming techniques, assisting with projects, and ensuring students are equipped to succeed.
- Graded projects and weekly homework assignments for 36 students, giving individually tailored feedback.

SAT Math Tutor | Capital Educators Jan 2024 –

- 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 weekly lessons and compiling biweekly diagnostic tests.