Alic Yao (Hongkun Yao)

hyao17@jh.edu | 412-888-9806 | https://www.linkedin.com/in/hongkun-yao-03bb63234/

EDUCATION

Johns Hopkins University **Baltimore, MD**

***Master of Science** in **Data Science** | **GPA**: 3.66/4.0* *08/2024 –05/2026 (expected)*

Relavant Coursework: Applied Statistics and Data Analysis, Introduction to Data Science, Introduction to convexity, Computing for Applied Mathematics, Mathematical Image Analysis, Artificial Intelligence, Computer Vision

University of Pittsburgh **Pittsburgh, PA**

***Bachelor of Science** in **Data Science**, **Minor** in **Computer Science** | **GPA**: 3.60/4.0* *08/2020 – 04/2024*

Relavant Coursework: Machine Learning, Intro to Nature Language Processing, Probability Theory, Linear Algebra

Academic Honor: Dean's List Award

PROFESSIONAL EXPERIENCE

Capgemini **Beijing, CN**

Data Analysis Intern *05/2023 – 07/2023*

- Developed **Python web scraping tool** using **Selenium**, allowing the Automated extraction and pre-processing of over **50,000** user data.
- Designed a streamlined pipeline to store data in **MongoDB, r**educing work time by **60%**.
- Leveraged the **Python wordcloud** library to create **dynamic visualization** reports, explored and unified the **BIO** label list for business analysis purposes, providing actionable insights to key business stakeholders.
- Conducted sentiment analysis on **1,000+** user's feedback and comments, applying **BIO** sequence labeling scheme manually for Named Entity Recognition (NER) preparation; improving user sentiment insights by **20%**.

PROJECTS

PPG Paint Prediction & Analysis *01/2024 – 04/2024*

- Conducted Exploratory Data Analysis using **R programming** to visualize the relationship between color models and crucial aspects of paint color, enabling a comprehensive understanding of the data distribution and key insights.
- Developed and fine-tuned 2 predictive models (**GLM**, **Bayesian linear models**) to predict paint properties and popularity, achieving **85%** accuracy in the prediction.
- Utilized **Logistic Regression** and **RF** for paint popularity prediction to learn patterns associated with top selling paint colors, enabled data-driven inventory management and marketing strategies, resulting in increased sales and customer satisfaction.

Database Development for ArborDB IoT Forest Monitoring System *08/2023 – 12/2023*

- Performed conceptual design using **Entity-Relationship (ER) diagrams**, constructed **PostgreSQL** table structures, and managed entity relationships.
- Created **B-tree indexes** on key fields (timestamp, area), improving query performance by **40%**.
- Implemented database connectivity and interaction via **Java JDBC**, ensuring stable data operations. Encapsulated data operations to complete the persistence of business logic, ensuring stable data operations for **1,000+** daily data transactions.

Diabetes Prediction *08/2023 – 12/2023*

- Built and validated predictive models using **R**, including regularized **logistic regression** and **Random Forest**, achieving an **92%** accuracy in diabetes prediction
- Identified key predictors of diabetes and found that men are more likely than women to develop diabetes, providing actionable insights for targeted interventions.

SKILLS

- **Certification**: Database and SQL for Data Science
- **Programming & Software**: Proficient in Python (NumPy, Pandas, Scikit-Learn, Matplotlib, Request, Selenium, Automating ETL pipelines), SQL, Excel, R, Power BI, Tableau, Jupyter Notebook, MongoDB, and AWS
- **Model Development**: Classification (KNN, SVM, Decision Tree, RF, QDA, LDA), Regression (Linear, Logistic), Clustering (K-mean, GMM, Embedding), PCA, Cross Validation