Weihang Gao

8550 Costa Verde Blvd, San Diego, CA 92122 (858)922-8311 wgao20@jh.edu

EDUCATION

MS, **Data Science**, *Johns Hopkins University* Aug 2021- Jan 2023 (expected)
BS, **Data Science** (GPA: 3.6 of 4) *University of California, San Diego* Aug 2017-June 2021

INTERNSHIP EXPERIENCES

Course Tutor, University of California - San Diego Jan 2019 - Apr 2019

Tutored core course in Data Science major using **Python**

- Tutored "Programming and Basic Data Structures for Data Science", a core course for undergraduate students majoring in data science (10 hours per week).
- Held lab hours to help students work on lecture materials and programming assignments.
- Graded student's assignments and exams on Grade-scope.

Database Engineer, Shenzhou General Data Technology Ltd. Jun 2018 - Sep 2018

*Manage Shentong Database by **SQL and Python***

- Wrote scripts to crawl articles in several online forums using **Python (Beautiful soup)**.
- Cleaned the text data and imported new data to a large-scale database using **SQL**. Managed the frame of the database and executed SQL queries.

SELECTED RESEARCHES & PROJECTS

A Research on US Government Spending (Machine Learning, Parallel Computing, Python) Sep 2020 - Current

- Used USAspending database (**110GB**) to predict the cost-overrun ratios for certain contracts.
- Applied **Parallel Computing** in data cleaning and model building using **DASK**, **Vaex**, and **RAPID**.
- Tried both **One-Hot Encoding** and **Label Encoding** as feature engineering. Selected the features that correlated with cost-overrun ratios by **Correlation Coefficient** and **Paired T-Test**.
- Trained fine-tuned multi-classification models such as **SVM**, **Logistics Regression, Ensemble Learning, XGboost** in **Sklearn**. Currently, XGboost achieved the best performance, with 78% test accuracy.

COVID-19 Geographical Analysis (Machine Learning, Natural Language Processing, Python) Sep 2020 - Jun 2021

- Built predictive models with multiple data sources, such as Google Mobility Report, CDC COVID Cases Report, Yelp reviews, Hospital record, PPE(Personal Protective Equipment) distribution, and articles scraped from news websites (CNN, CNBC).
- Got numerical data from the news articles by applying **Topic Modeling** and **Sentiment Analysis**.
- Trained fine-tuned prediction models such as **Linear Regression**, **Random Forest**, **Long Short Term Memory**.
- Took the advantage of UCSD Cloud computing resources. Created a **Docker Repository** to set up the environment.
- Connected with the ***CA Notify*** team and worked on COVID-19 prediction and analysis.

PC User Persona Analysis (Machine Learning, Python, C++, SQL, Intel SUR) Sep 2020 - Mar 2021

- Collected PC user data by applying the INTEL System Usage Report (SUR). Based on the user data such as RAM, OS, cpu_vendor, predicted user personas such as web user, casual user, gamer.
- Created an input library of INTEL SUR in **Visual Studio** using **C++**. Executed the input library to collect certain user data and saved the output as a database. Cleaned the raw data using **SQL** and **Python**.
- Applied **Hypothesis Test** and **Chi-square Test** to determine and validate the correlations between the user data and user personas.
- Trained fine-tuned multi-classification models such as **KNN**, **SGD**, **Logistic Regression**, **Random Forest, Multilayer perceptron** in **Sklearn**. The multilayer perceptron model has the best performance, with 86% test accuracy and 82% test F1 score.

Podcast Pro (Natural Language Processing, Python, AWS) Jan 2020 - Jun 2020

- Built an outline tool to automatically produce key words for each lecture podcast (only in UCSD).
- Served as the **Tech Lead** of the project team.
- Wrote scripts to download the podcast videos and audios by using **Chrome driver** and **Beautiful soup**. Uploaded the videos and audios to **AWS S3** Storage. Applied **AWS Transcribe** on the audios and downloaded the transcripts. Applied **AWS Rekognition** on the screenshots of the videos to get the content of lecture slides.
- Based on the transcripts and slides content, produced the key words for each lecture and each course by using **TF-IDF** and **Rapid Automatic Keyword Extraction (RAKE)**.

SKILLS

- **Technical Skills:** Python(Numpy, Pandas, Scikit-learn, Tensorflow, Matplotlib, Seaborn, DASK, Beautiful Soup), SQL, Java, AWS, R, HTML, Javascript, CSS
- **Software:** Microsoft Office Suite, Tableau, MATLAB
- **Coursework:** Predictive Analysis, Data Visualization, Analytics for Big Data, Parallel Computing, Data Mining, Data Structure, A/B Testing, Database, Text Analytics, Deep Learning, Probability Theory, Optimization, Simulation