QUOC-HUY NGUYEN
(657) 728-6866 | quochuyn1999@gmail.com | linkedin.com/in/quochuyn | quochuyn.github.io

EXPERIENCE

Data Scientist | GenAI | AWS | OpenAI | Utilities Jun 2024 - Present
SoCalGas

- Leading the deployment, maintenance, and enhancement of a machine learning risk model that processes >1M records annually. The model mitigates the impact of 600+ gas pipeline damages per year, resulting in >$500k savings while minimizing risks of injury or loss of life.
- Leveraged large language models (GPT-4o) and prompt engineering to identify and categorize root causes for >140k customer complaints, enabling systemic business process improvements and enhancing customer service. Implemented processing pipeline in a scalable, multiprocessing framework, accelerating processing time by ~300%.
- Automated monthly reports, reducing manual labor hours by approximately 98% and streamlining business workflows for better operational efficiency.

Data Scientist (Freelancer) | Keras | Tensorflow | Consulting | Healthcare May 2022 - Jun 2024
Upwork

- Finetuned a Keras EfficientNetV2 deep learning architecture on noisy EEG spectrogram data to identify seizures and other harmful brain activity patterns in critically ill patients, improving predictions by 9% and enhancing patient outcomes.
- Evaluated an overfitting ResNet-based TensorFlow convolutional neural network (CNN) on brain tumor imaging, diagnosing poor generalization and recommending regularization techniques to meet client requirements for medical research.
- Developed scripts using GeoPandas to identify regional and temporal attack patterns in malicious hacking attempts, pinpointing ~10 key geographic locations for further investigation and targeted cybersecurity measures.

Machine Learning Researcher | LLMs | AWS | Streamlit | Youtube May 2023 - Apr 2024
University of Michigan, School of Information

- Developed an end-to-end solution integrating the Youtube API, a website for real-time viewing, and an NLP-based (BERT or ChatGPT) filtering system, providing an intuitive interface for users to personalize and refine their video recommendations.
    - Configured an AWS RDS PostgreSQL database for user preferences that integrated with the Streamlit web app.
- Created a machine learning model that predicted NHL draft positions for prospective hockey players and clustered players together based solely on text report word embeddings, showing the potential of nontraditional methods into sports analytics.

Data Analyst | ETL | Fintech | Customer Sep 2021 - Apr 2022
Curacao

- Engineered a data pipeline for our mobile application product, enabling the extraction, transformation, and loading (ETL) of 16+ GB of Open Banking data from FinTech APIs to underwrite personal loans, identifying key credit risk levels to enhance credit decisioning.
- Leveraged customer data analytics to guide data-driven decision making, identifying high-volume user segments and optimizing physical space utilization by ~5% for 14 department stores.
- Optimized SAS code logic for routine reports, reducing processing time by ~10%.
- Acquired proficiency in SQL and SAS within 2 months, leveraging these tools to execute data analytic tasks.

Undergraduate Researcher | Pytorch | Matlab | Single-cell Data Science Jul 2019 - Mar 2021
Atwood Lab

- Implemented a Pytorch graph-based convolutional neural network (CNN) for semi-supervised classification model, based on a highly cited research article, to enhance cancerous cell identification accuracy. By leveraging innovative tools outside industry standards, such as principal graphs and graph neural networks, novel approaches were contributed to cancer research.

EDUCATION

Master of Applied Data Science, University of Michigan, Ann Arbor

- Awarded UMSI Millers Scholars Fund

Bachelor of Science in Mathematics, University of California, Irvine

- Data Science Specialization, Statistics Minor, Phi Beta Kappa, Summa Cum Laude

SKILLS

Programming: Python, SQL, PostgreSQL, R, SAS, C++, Matlab, Bash, Shell, Unix/Linux Terminal
Libraries: Numpy, Pandas, Scikit-learn, PyTorch, TensorFlow, Keras, PySpark, Statsmodels, NLTK, Matplotlib, Seaborn, Altair, Geopandas, Streamlit, ArcPy, Python-Docx, boto3
Datasets: Utilities, Customer, Financial, B2B, Healthcare/Clinical, Geospatial, Causal Inference, JSON, Experiments, Web Scraping
Tools: Git, DVC, AWS (S3, EC2, RDS, Sagemaker), Azure, Power BI, Domo, SAP HANA Studio, LaTeX, Agile Methodology