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

EXPERIENCE

Data Scientist | Pandas | ArcPy | SQL | Utilities Jun 2024 - Present
SoCalGas

- Leading the development, management, and enhancement of a machine learning risk model that handles 1M+ records annually, mitigating the risk of gas pipeline damages that can lead to injury or loss of life.
- Automated the creation of routine document mailers by developing a Python script, reducing manual work from approximately 15 hours/month to 15 minutes/month.

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

- Utilized Keras to finetune an EfficientNetV2 deep learning architecture on EEG spectrogram data to identify seizures and other harmful brain activity patterns in critically ill patients for better healthcare outcomes, improving predictions by 9%.
- Delivered expert consulting on Python programming: data engineering, code reviews, project planning, and documentation.
- Evaluated an overfitting ResNet-based TensorFlow convolutional neural network (CNN) on brain tumor imaging aligning it with the requirements of a client's medical research within the healthcare domain.
- Developed Python scripts using GeoPandas to understand patterns and trends in malicious hacking attempts, identifying ~10 key geographic locations for further investigation.
- Implemented SQL queries to extract valuable insights from customer billing transactions that were presented to a stakeholder.

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

- Developed an end-to-end solution that integrated with the Youtube API and a website for real-time content viewing and NLP-based (BERT or ChatGPT) filtering to give users more control over their Youtube 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 | Pandas | Scikit-learn | ETL | SQL | SAS | Microsoft Excel | 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.
- Designed and built a scikit-learn random forest classification model that surpassed a 70% prediction accuracy, discerning key thresholds of diverse credit risk levels.
- Leveraged customer data analytics to guide data-driven decision making, identifying high-volume user segments and optimizing physical space utilization.
- Optimized reports in SAS/Excel by removing unnecessary joins, functions, and references, leading to ~10% time reduction.
- 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 from a highly cited research article for identifying cancerous cells.
- Researched skin cancer single-cell data using methods for principal graphs and graph neural networks.
- Performed computational biology research at MathBioU & Math EXPLR, a 6-week summer research program.
- Collaborated as a co-author in a research report to present findings to a diverse >15 person audience.

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 (4 years), SQL (2 years), PostgreSQL, R, SAS, C++, Matlab, Bash, Shell, Unix/Linux Terminal
Libraries: Numpy, Pandas, Scikit-learn, PyTorch, TensorFlow, Keras, Hugging Face, PySpark, Statsmodels, NLTK, Matplotlib, Seaborn, Altair, Geopandas, Psycopg2, Streamlit, ArcPy, Python-Docx
Datasets: Customer, Financial, B2B, Healthcare/Clinical, Geospatial, Causal Inference, JSON, Experiments, Web Scraping
Tools: Version Control (Git, Github, DVC), Azure DevOps, Jupyter Notebook, Business Intelligence (Power BI, Domo), Microsoft Excel, Cloud Infrastructure (AWS S3/EC2/RDS/Sagemaker), DBeaver, RStudio, LaTeX