SCOTT PENCO

Machine Learning and Data Science Devotee Interested in Making an Impact With Designing Data Driven Solutions Using Mathematical Models and MLOPS. Open to Collaborating With Others and Learning From Professionals.

Montreal, CANADA | scott.penco@alumni.utoronto.ca | LinkedIn | 647-962-9436

EDUCATION

MIT Institute for Data, Systems, and Society (IDSS) - Data Science and Machine Learning

Applied AI/Data Science and Mentorship, Data Analysis: Statistical Modelling and Computation in Application, Foundations of Statistics, Statistical Modelling, Probability , Machine Learning.

Harvard Extension School- Graduate Certificate in Bioinformatics

Computer Science, Bioinformatics, Data Structures and Algorithms

University of Toronto - HBSc Double Major in Biology and Neuropsychology (Joint Hons.)

Fundamentals of Statistics, Calculus, Research Design Analysis, Insight Into Neural Network Models With Applications to AI

PROJECTS

Malaria Capstone Project - *MIT IDSS* GITHub Repo

- Presented a Malaria Classification Capstone Project to the MIT IDSS. Model was designed using CNNs.
- Model was built using Convolutional Neural Networks built on a 30,000 image dataset. Preprocessing and cleaning of image data was performed. CNN was built using the library Keras, and compared to a pre-trained CNN (VGG16).
- Chosen model out performed VGG16 by .03 in precision and recall, having an F-1 score of .98.
- SHAP values were calculated for the chosen.

ML Prediction GITHub Repo

- A Linear model was used to create a predictive machine learning model for Boston house prices
- Data was explored using univariate and bivariate analysis.
- The model equation/ coefficients were determined using OLS, these highlighted the important features for determining housing pricing in Boston.
- Features that were not statistically significant were dropped, resulting in an increase in R^2 values by .001

Data Analysis and Visualization GITHub Repo

- Application of dimensionality reduction on the auto-mpg dataset, to ensure costumers queries are answered more quickly.
- PCA, t-sne and clustering was used to reduce and visualize the data into lower dimensions to extract insights.
- 3 types of groups were determined.

RAG and Fine-tuning of LLMs GITHub Repo

- Academic papers were scraped and vectorized for Retrieval Augmented Generation on a Large Language Model (Open AI)
- Prompts and answers from the same academia were generated and used to fine-tune and change the models layers to train to improve its proficiency in answering questions.
- Models were compared and evaluated based on how well they answered potential questions in the field of '*Machine Learning for drug development.*'

SKILLS

Data Science & Programming: Python, Numpy, Pandas, sklearn, MatPlotLib, Seaborn, Keras, PyTorch, SQL

Other Certifications: Foundations in Data Science (University of Waterloo), Data-Camp Data Science Certification (concurrent), Public Speaking and Vocal Training (Speech Science).

Languages:

- Native English proficiency, B2 proficiency in French