Hao Dinh Sy

dinhsyhao@gmail.com
(84) 987 606 428
Hanoi, VN
LinkedIn: dinhsyhao

Data Scientist

Strong personal self-learning ability, serious work, careful, organized, quick thinking, accepting challenges to new knowledge Sincere ability, treat people sincerely, have good teamwork and organization, and work actively In the face of different things, have strong opinions, have time to witness, wise independent thinking, excellent expression For the strong interest in the IT industry, and continue to study hard to improve their technical level Code habits, clear requirements, naming conventions, and well-structured logic.

WORK EXPERIENCE

Viettel Network

Data Scientist

Jul 2022 - now

- Gathered and analyzed requirements from stakeholders to understand project objectives and expectations.
- Acted as a liaison between the customer and the data science team, translating business needs into actionable technical specifications.
- Implemented data science tools, such as Jupyter Notebook/Lab/Hub, MLflow, and Code-Server, to establish a foundational working environment for team members.
- Implemented Spark 3 for efficient handling of big data clusters, and devised guidelines for a streamlined pipeline to execute jobs using PySpark.
- Collaborated on data science projects, contributing to the development of end-to-end solutions.

Samsung Vietnam Mobile R&D Center

Data Analyst

Jun 2021 - Jul 2022

- Processed and cleaned raw machine log data to ensure its suitability for input into machine learning models.
- Built & Trained the Model using various algorithm: Random Forest Classification, XGBClassification,... and analyzed various scores: accuracy, f1-score, Confusion Matrix, AUC Score.
- Fine-tuned the Model using Optuna for minimizing the loss function: RMSE, MAE, Mixout Loss.
- Packaged the full pipeline and deploy to MLFlow.

PROJECTS

Predict False Alarm for manufacturing machines.

- The manufacturing machine test kpis of a device if it is **Fail** or not, however sometimes the testing process give wrong results (called False Alarm). This Project is to create a ML-based model which feeds on the historical logged data to predict these above.
- Achieved an **86%** accuracy during the testing phase, significantly reducing the need for retesting efforts.

NBME - Score Clinical Patient Notes (Kaggle competition).

- Automatically identified keywords (clinical case) that match in student notes for scoring system.
- Got top **8%** with the competition micro-averaged f1 score of 0.888 compared to the 1st score of 0.894.
- Performed Masked Language Modeling (MLM) on the training data for model generlisation.
- Applied the Cross Validation (CV) of 5-fold on the training data for optimizing data usage and prevent overfitting.
- Trained various state-of-the-art NLP models: DeBERTa, BERT, CoCo-lm,... using **Transformers** Trainer API, and custom training using PyTorch only.
- Trained the model on unlabeled data with **Pseudo Labeling** method.
- Experimented with the CV score of various model and do the model ensembling by using **Optuna** on the CV score.
- Tried the **8bit optimizer** to train large model like: **DeBERTa-v2-xlarge** on Kaggle limited environment.

Counting number of base station cells & Identified low radio signal regions

- Utilized cluster algorithms based on geolocation data from end devices (mobile phones) to automatically detect regions that meet specified criteria.
- Developed a DBSCAN model with optimized parameters that surpassed previous solutions, achieving a 15% increase in the number of correctly identified clusters.

Forecasting number of customers using service

- Utilized users' historical activities to forecast the number of users utilizing the service in the upcoming time interval. This predictive analysis informs decisions on selectively turning off cells for electricity conservation.
- Implemented an LSTM model using the most recent week's data, yielding a result with a Mean Absolute Error (MAE) of approximately 5 users difference.

SKILLS

- Programming: Python, C++, bash script.
- Supervised Learning: Regression & Classification model, tree based model, CatBoost, LightGBM, AutoML.

- Unsupervised Learning: Kmeans, DBSCAN, HDBSCAN, hierarchical clustering.
- Semi-supervised Learning: self-train, contrastive learning.
- Reinforcement Learning: Offline RL, Model-free algorithm.
- Big Data: HDFS, Spark.
- Deep Learning: PyTorch, sequence model.
- Data Visualization: Tableau.
- AWS: EC2

EDUCATION

B.S. Information Technology

VNU University of Engineering and Technology

Jul 2017 – Jul 2021

- Course Recommendation: Developed a recommendation system utilizing student scores to suggest optimal next-term courses, aiming to enhance academic performance.
- Question-Answering for educational documents: The medical students have many lectures to remember. This project is to create a question-answering system using GPT-2 model from that documents.
- Cell Segmentation: Segment single cell images from a multi-cell images using unsupervised learning technique.

M.S. Computer Science

VNU University of Engineering and Technology

Nov 2022 – now

- Research topics: Large Sequence Models for Sequential Decision-Making, Reinforcement Learning, Transformer.

CERTIFICATION

- TOEIC: 750
- Kaggle Expert
- Professional Data structure & Algorithm Certificate issued by Samsung