Ha Dang Vu

DATA SCIENCE INTERN — MACHINE LEARNING — STATISTICS

☎ 647-551-3275 ✉ dangvuha.2803@gmail.com linkedin.com/in/hadangvu github.com/hdangvu

Education

HEC Montréal **Sep 2025 – Present**
Master of Data Science and Business Analytics *Montréal, Québec*

University of Waterloo **Sep 2020 – Dec 2024**
Bachelor of Mathematics in Computational Mathematics *Waterloo, Ontario*

- Minor in Computing, Combinatorics and Optimization
- **Best Insight** Award at ASA DataFest 2024

Vector Institute **Jul 2024**
CIFAR Deep Learning Reinforcement Learning 2024 – Summer School *Toronto, Ontario*

Technical Skills

Languages: Python, R, SQL, MATLAB. **ML & Data**: NumPy, Pandas, Scikit-learn, PyTorch, TensorFlow, HuggingFace. **Tools**: Git, Linux, Jupyter, Power BI, Tableau.

Projects

BIXI Montréal Trip Behavior Analysis | *R, RMarkdown, GLM, Logistic Regression*

- Modeled **trip duration** and **rush-hour usage** using interpretable **log-linear** and **binomial GLMs**, incorporating temporal and weather effects.
- Found weekend trips ∼**10% longer**, rush-hour odds ∼**50% lower** on weekends, and **heavy rain reducing rush-hour activity** on Fridays/Saturdays; identified strong **station-level heterogeneity** motivating mixed models.
- Built a reproducible **data preparation pipeline** to clean, validate, and aggregate BIXI trip and weather data into analysis-ready tables.

Efficient Financial Sentiment Modeling via Knowledge Distillation | *Python, PyTorch, HuggingFace Transformers*

- Designed a lightweight financial text classification model using **Knowledge Distillation** to transfer knowledge from a large model to a smaller one, reducing model size by **10×** (109.5M → 11.7M) with minimal performance loss.
- Achieved **97.35% test accuracy** and **0.9626 macro-F1** on real financial news data, with **no increase in prediction time** (∼2.0 ms per document), making the model suitable for large-scale or real-time use.
- Built an end-to-end **financial news data pipeline**, including web scraping, text cleaning, schema design, and structured storage (raw/processed/sample layers) to support scalable ML training.

MNAR Sensitivity Analysis for Predictive Modeling | *R, Monte Carlo Simulation, Logistic Regression, Missing Data*

- Built a **Monte Carlo framework** to analyze **MNAR effects**, separating degradation by **feature importance** and **missingness intensity** (25%–70%).
- Showed **PMM failure under MNAR** (up to **11% accuracy loss**) and that **delta adjustment** recovers ∼**87%** of signal, while **mis-specified corrections** amplify bias.
- Created a **reproducible data pipeline** to harmonize schemas across **real and synthetic datasets**, enabling scalable **Monte Carlo simulation** under controlled MNAR mechanisms.

Experience

Research Assistant **Nov 2025 – Present**
HEC Montréal *Montréal, Québec*

- Performed large-scale analysis of **Machine Learning applications in Information Systems**, systematically extracting structured features (data types, algorithms, contexts) from academic studies using **Covidence**.
- Built a reproducible review pipeline using **Covidence**, applying consistent inclusion criteria and data schemas - mirroring real-world **data curation and feature engineering** workflows.

Undergraduate Research Assistant **Sep 2024 – Dec 2024**
WiM Directed Reading Program @ UWaterloo *Waterloo, Ontario*

- Applied **convex optimization** methods (PGD, FISTA, ADMM) to **image denoising** and **deblurring tasks**, modeling sparse signal recovery via L_1-regularization.
- Implemented and evaluated **image restoration** algorithms in MATLAB, comparing convergence, stability, and reconstruction quality across iterative solvers.