John Matthew Van

johnvan1999@gmail.com | linkedin.com/in/johnvan | https://johnvan1.github.io/

Education

UC Berkeley School of Information | GPA: 3.9/4.0
Master of Information and Data Science May 2024

- Relevant Coursework: Data Science Programming, Research Design, Fundamentals of Data Engineering, Applied Machine Learning, Natural Language Processing, Computer Vision, Machine Learning at Scale

University of California, Berkeley
Data Science, B.A. & Economics, B.A. (Double Major) May 2021

Skills & Tools

Programming: Python (primary), SQL, R, Java

Machine Learning: Scikit-learn, PyTorch, TensorFlow, HuggingFace Transformers, BERT, T5, LLMs (Claude, GPT)

Data Platforms: Google Cloud Platform, BigQuery, Snowflake, Databricks, AWS, Spark, Hadoop

Experimentation & Statistics: A/B Testing, Causal Inference, Hypothesis Testing, Regression Modeling

Work Experience

CVS Health | Woonsocket, RI
Data Scientist July 2024 – Present

- Engineered Smart Health Journey (SHJ) Medicare, a multi-model health risk product that identifies high-risk patients across 3M+ Medicare members, driving $8M+ in revenue by reducing leakage and improving case mix
- Built & deployed XGBoost & CatBoost health risk models for readmission and new-member risk prediction, achieving an AUC of 0.76; model scores actively used for 3M+ Medicare members daily
- Conducted comparative analysis revealing the IP6 model outperforms New Member risk model after month 2, contradicting core modeling assumptions; presented findings to VP of Clinical Analytics, driving org-wide model adoption across 3M+ Medicare members, increasing medical cost savings by $2M/year
- Designed and shipped an automated statistical validation pipeline for SHJ Medicaid, aggregating 5 statistical output checks (distribution drift, score stability, z-score checks) to gate production deployments and eliminate manual QA
- Built a receptivity model predicting Medicaid member engagement, achieving 14% higher response rate vs. baseline

CVS Health | Woonsocket, RI
Data Science & Analytics Intern June 2023 – Aug 2023

- Built a multi-linear regression model in Python to score customer retention risk, achieving 82% predictive accuracy
- Queried & engineered features from 300M+ rows of transactional data via SQL, identifying 25 final key features

Infosys Limited | Los Angeles, CA
Software Engineer June 2021 – Jan 2023

- Architected Python & Tableau analytics dashboards, enabling KPI visibility for 4 distinct enterprise clients
- Led end-to-end migration of a ~25TB legacy SQL database to Snowflake, cutting infrastructure costs by $50k/year

Projects

Fraudulent Job Postings

- Fine-tuned BERT, Bag of Words, TF-IDF, & GloVe embeddings for fake job detection; TF-IDF achieved F1 of 0.91
- Resolved class imbalance via dataset rebalancing/BERT class-weight tuning, optimizing for precision-recall tradeoff

DeeBoT

- Engineered GPT-4 into a personalized DBT therapy chatbot through custom system prompt design, synthetic conversation generation and domain-specific fine-tuning on core DBT skills
- Designed & deployed a rigorous evaluation framework scoring 81 synthetic conversations across 23 DBT adherence criteria, executing 9,315 GPT-4 evaluation calls and achieving 70% DBT standard adherence

Interests

Swimming | Junior Olympian, Swim Captain

- Ranked Top 30 in California high school swim, qualifying to the 2017 State Championship in Fresno, CA