VARUN VADDI

varunvaddi30600@gmail.com | (713) 539-6996 | www.linkedin.com/in/varunvaddi/ | https://varunvaddi.github.io/

WORK EXPERIENCE

Data Science Intern, Wild Genomics, CA *May 2025 - Aug 2025*

- Resolved low-quality sequencing data that limited ML insights by engineering an end-to-end **Python pipeline** with adaptive filtering & QC, increasing **analysis-ready data from 30% to 80%** and enabling more accurate downstream predictive modeling.
- Performed **similarity-based clustering** of sequences at 97-99% identity to **reduce redundant data by ~90%,** improving feature reliability and cross-sample consistency for ML workflows.
- Applied **threshold-optimized similarity classification** (BLAST, 95% identity), increasing label **accuracy from 85% to 95%**, enhancing granularity from family to genus level, and enabling precise supervised ML training.

Data Engineer, University of Houston, TX *Jul 2024 - Dec 2025*

- Eliminated **70% of manual admissions** data processing by architecting automated **Data pipelines** integrating multiple sources, enabling **real-time sync of 40K+ application** records per cycle and accelerating admit decisions.
- Resolved fragmented data across application and test portals by building **unified extraction workflows**, improving analytics accuracy and providing consistent, **model-ready data** for downstream predictive analyses.

Software Engineer, Accenture, Hyderabad, India *Sep 2021 - Dec 2023*

- Optimized large-scale customer data systems managing **80M+ records** to address slow analytics and fragmented insights; streamlined workflows and improved **deployment speed by 30%**, enabling faster, data-driven sales decisions.
- Led large-scale **CRM data migrations** and developed analytics pipelines using **SQL**, enabling detailed conversion analysis and funnel optimization that contributed to **$25M in revenue growth.**
- Defined key performance metrics and automated **Tableau** dashboards and reports to monitor feature impact and closed-won deals, enabling actionable insights and increasing **conversion rates by 10%.**

PROJECTS

NYC Taxi Data Pipeline

- Engineered an end-to-end analytics pipeline **(Snowflake, dbt, Airflow, Kafka)** to address slow ML feature preparation on **8.6M+ NYC taxi trips**, reducing feature-ready **data prep time by 95%** with **37 automated validation checks** and enabling real-time demand modeling.
- Analyzed **$180M+ quarterly revenue** using **SQL** and **Tableau** to uncover **30% peak-hour** demand lift and **18% higher credit-card** revenue, informing feature selection, edge-case handling, and downstream predictive models.

Google Ad Policy Compliance with RAG

- Developed a hybrid **RAG** system **(BM25, BGE embeddings, FAISS, reranking)** over 341 policy chunks to address slow manual ad reviews (5-10 min per ad), reducing review **latency to <1 second per ad** while achieving **80% classification accuracy** and **78% Recall@5**, enabling high-throughput compliance at scale.
- Implemented structured LLM outputs **(Gemini)** with citations and confidence scores to resolve inconsistent ad policy decisions, eliminating hallucinations and enabling real-time **classification of 1,000+ ads/sec** with **95% policy coverage.**

Named Entity Recognition on X posts

- Designed a **BERT-based NER** pipeline with **SpaCy** preprocessing and IOB tagging to address inconsistent entity extraction from **11,000 X posts**, improving **macro F1 to 89%** across 7 entity types and enabling fast, accurate trend monitoring.
- Streamlined sub-word label reconciliation using **majority-vote and first-token** rules to reduce **labeling errors by 12%,** enabling precise and scalable downstream **social media analytics** for social media trends.

EDUCATION

University of Houston **TX, USA**

Master of Science (MS), Engineering Data Science | ***GPA: 3.7/4*** *Dec 2025*

SKILLS

Programming Languages: Python, SQL, R, Java, Linux.

Tools/Frameworks: Tableau, Power BI, Advanced Excel, Pandas, NumPy, Matplotlib, Seaborn, Plotly, PyTorch, TensorFlow, Keras, Scikit-learn, Snowflake, dbt, Airflow, Docker, Kafka, Git, CI/CD, Kubernetes, Jira.

Competencies: ML, Deep Learning, Neural Networks, Transformers, NLP, Generative AI & LLMs, RAG, Predictive Modeling.