VISHAL REDDY DEVIREDDY

◦ Raleigh, NC 27606 ◦ vdevire2@ncsu.edu ◦ +1(704)747-8151 ◦ LinkedIn ◦ GitHub

EDUCATION

North Carolina State University - *Master of Computer Science*-3.88/4 CGPA **Aug 2024- May 2026**
Coursework: Software Engineering, Object-Oriented Design and Development, Automated Learning and Data Analysis, Human Computer Interaction, Computer Networks, Neural Networks.

SRM Institute of Science and Technology - *Bachelor of Computer Science*-8.59/10 CGPA **Aug 2020- May 2024**
Coursework: Database Management system, Artificial Intelligence, Machine Learning, Data Structures and Algorithm, Object-Oriented programming.

TECHNICAL SKILLS

- **Programming:** Python, C++, C, Ruby, JavaScript.
- **Web Technologies:** HTML, CSS, GIT, React JS, Angular, Vue, Rest APIs.
- **Machine Learning:** Scikit-learn, TensorFlow, PyTorch, XGBoost, Transformers, BERT
- **NLP & LLMs:** RAG, Embeddings, Tokenization, Text Classification, LangChain
- **Computer Vision:** CNNs, YOLO, Image Classification, Object Detection, OpenCV.
- **Data Science:** Pandas, NumPy, Matplotlib, Seaborn, Power BI, Excel, Tableau.
- **Cloud:** AWS - SageMaker, Lambda, S3, GCP – Vertex AI, Azure - Databricks.
- **Databases and Operating Systems:** MySQL, PostgreSQL, SQL, Oracle, Windows, Ubuntu, Linux.
- **MLOps:** MLflow, Model Deployment, CI/CD Basics, Vector Search (FAISS), Monitoring & Model Evaluation
- **Certifications**: Machine learning , AWS Cloud9 , Oracle SQL, Azura AI Fundamentals.

EXPERIENCE

Dev Systems - Data Science Intern Vijayawada, India **Jan 2024 - July 2024**

- Queried and analyzed 5M+ customer records using SQL and built Python-based cohort, retention, and funnel analysis pipelines that revealed key behavioral trends and supported data-driven decision-making across teams.
- Designed rich EDA dashboards and an interactive Gradio ML demo suite, enabling 20+ stakeholders to visualize patterns, validate model outputs, and significantly accelerate model review and iteration speed.

PROJECTS

Privacy-Preserving Local Search & Messaging Assistant (LLM + RAG): Aug 2025 - Nov 2025

Tech Stack: Python, LangChain, FAISS, SentenceTransformers, FastAPI, SearxNG, NumPy, Pandas, JSON, REST APIs

- Developed a Retrieval-Augmented Generation (RAG) system indexing over 50k documents with FAISS, enabling fast, private semantic search without relying on cloud-based APIs and ensuring complete user data confidentiality.
- Designed an end-to-end data pipeline with embedding generation, chunking, ranking, and analytics, including cluster visualizations and hit-rate evaluation plots to assess retrieval quality and model performance.
- Integrated SearxNG and local LLM inference through FastAPI services, improving answer relevance by 23% using optimized chunk sizes, embedding models, and prompt-engineering techniques.

Customer Churn Prediction in Telecommunication Using Machine Learning Algorithms: Jan 2024 - May 2024

Tech Stack: Python, Scikit-learn, TensorFlow, Pandas, NumPy, Matplotlib, Seaborn, Jupyter Notebook, SQL

- Processed and engineered 7k+ customer records by performing extensive EDA with heatmaps, churn-distribution plots, feature correlations, and data-cleaning techniques to improve model reliability and interpretability.
- Implemented and compared Logistic Regression, Random Forest, KNN, and Neural Networks using cross-validation, achieving 85.6% accuracy while evaluating models through ROC-AUC, confusion matrices, and F1 scores.
- Generated actionable insights using feature-importance visualization and business metric analysis, identifying top churn drivers that support strategic retention planning and customer engagement improvements.

YOLOv8 Object Detection System Aug 2024 - Nov 2024

Tech Stack: MySQL, DBMS, Flask, Python, HTML, Rest APIs and CSS.

- Trained a YOLOv8 model with augmentation (rotations, contrast shifts, scaling) to improve robustness and accuracy.
- Evaluated detection performance using mAP and precision–recall curves to analyze model quality across classes.
- Deployed the model via a FastAPI endpoint for real-time inference, added visualization outputs, and built an interactive testing interface that supports live predictions, model monitoring, and rapid debugging.

PAPER PUBLICATION

Publication Title: *"Anticipating Customer Churn in Telecommunication using Machine Learning Algorithms for Customer Retention"*, co-authored with Rajarajan K. and Priya S., presented at the **2024 Third International Conference on Distributed Computing and Electrical Circuits and Electronics (ICDCECE)**.