Krishna Thakar

krsnathkr.github.io | (858)-408-5293
krsnathkr@gmail.com | LinkedIn | GitHub

Technical Skills:

- **Programming Languages**: Python, R, Java, SQL (PostgreSQL, MySQL), MongoDB
- **Frameworks & Libraries**: Pandas, NumPy, Scikit-learn, Matplotlib, Seaborn, Plotly, Streamlit, OpenCV, NLTK, Hugging Face Transformers, TensorFlow, PyTorch, Keras, CUDA, LSTM, CNN, RAG LLM
- **Tools**: Snowflake, Databricks, Hadoop, AWS (EC2, S3), Docker, Kubernetes, CI/CD pipelines, Scala, Apache Spark, Apache Beam.

Education:

- Undergraduate in Computer Science w/ Minor: Data Science – 3.9 GPA — May 25'
  Southeast Missouri State University
  Dean's List x 6 | President's List x 4
- Coursera Certifications:
  Supervised Machine Learning: Regression and Classification - Deeplearning.AI <Link> — Feb 24'
  Exploratory Data Analysis for Machine Learning – IBM <Link> — June 24'

Job Experience:

Undergraduate Student Researcher - Southeast State Missouri University <Research Link> — Jan 25' – April 25'

- Worked under the mentorship of Dr. Mohamed Abu Sheha and Dr. Emmanuel Thompson, comparing traditional ML models and deep learning models against fine-tuned RoBERTa for three-way sentiment classification.
- Processed ~7 million Yelp reviews, performing **text cleaning, tokenization, lemmatization, negation handling, vectorization**, and balancing sentiment classes into equal thirds.
- Developed and evaluated multiple models - including Logistic Regression, SVM, Naïve Bayes, Random Forest, BiLSTM, LSTM, CNN, RNN, GRU, and a fine tuned RoBERTa - **using stratified cross validation, confusion matrices, and ROC curves** to measure accuracy, precision, recall, F1, and ROC AUC.
- Demonstrated that RoBERTa achieved top performance (accuracy 0.80, AUC 0.93) through systematic cross-validation and in-depth analysis.

Information Technology Staff – Southeast State Missouri University — Sept 23' – Jan 25'

- Provided front-line technical support to 100+ students and faculty, resolving issues with software, printing, and system functionality at an on-campus IT help desk.

Projects:

ASA DataFest 2025 – Winner, Best Use of Statistical Analysis | R, RStudio: <Link> — April 25'

- Analyzed 194,000+ U.S. commercial lease records to predict whether technology-sector tenants would relocate or renew.
- Improved model precision by 18% by cleaning duplicates, engineering variables like Space Type, Lease Year, and scaling features to align distributions before training **Logistic Regression, Classification Tree, and Random Forest** models.
- Focused on deriving insights through structured EDA, developing predictive models with balanced error metrics, and interpreting feature importance for real-world business decision support.
- Delivered an end-to-end analysis from raw data preprocessing to model evaluation within 24 hours, reinforcing data storytelling and hypothesis-driven modeling under tight timelines.

Sales & Customer Data Pipeline | Azure Data Factory, Databricks, dbt, Apache Spark, ADLS Gen2: <Link> — May 25'

- Engineered a sales & customer data pipeline using dbt and Spark with a **Medallion architecture**, reducing analytics reporting time by 40% and processing over 1M+ records.
- Developed PySpark transformations in Azure Databricks to cleanse, dedupe, and enrich data across bronze, silver, and gold layers, delivering sub-5-minute SLA processing and reducing compute costs by 30%.
- Modeled and versioned 15+ data assets with dbt on Databricks - incorporating snapshots, automated tests, and lineage documentation - to support reliable, self-service BI reporting.

Knowbl – Agentic RAG Assistant with Real-Time Web + PDF | LangChain, MCP, BeautifulSoup: <Link> — April 25'

- Built an **agentic RAG pipeline** combining real-time web search and document Q&A, using FAISS vector indexing and OpenAI embeddings to return citation-backed answers; handled 10+ queries in local tests with consistent <2s response time.
- Processed and embedded unstructured data (web articles, PDFs) into a **FAISS vectorstore** using recursive chunking (10K tokens w/ 500 overlap), improving document retrieval relevance for question answering.
- Built a **modular backend with FastMCP** to support advanced filtering (e.g., domain restrictions, text inclusion, date limits), allowing targeted analysis across web-sourced data.