SHREYAS KULKARNI

+1-669-977-9227 | skulkarni.sk.18@gmail.com | LinkedIn | Github | Publication | Certifications

Education

San Jose State University Aug. 2023 – May 2025
Master of Science in Computer Engineering *San Jose, CA*
Coursework: Data Structures and Algorithms using C++, Machine Learning, Data Mining, Deep Learning, NLP, OS

Technical Skills

Languages: Python, C/C++, SQL (Postgres), Shell Scripting, MATLAB
Tools: Tensorflow, PyTorch, OpenVINO, ONNX, AutoML, MLFlow, Pandas, Numpy, Scikit-Learn, Hugging Face, Git, GitLab, Docker, Airflow, Kubernetes, Google Cloud, AWS EC2, S3, Databricks, Hadoop, PySpark, Jupyter Notebook

Experience

Deep Learning Project Assistant Aug. 2024 – Present
San Jose State University *San Jose, CA*

- Utilizing zero-shot (all-MiniLM-L6-v2) and pre-trained generative models (BERT, Sentence Transformers, **Llama 3.1**) using PyTorch to boost resume parsing and candidate job matching accuracy, achieving a 95% efficiency
- Implementing Nvidia embedding models and **Milvus** vector database for efficient query retrieval, making the application scalable and context-aware AI solution relevant to conversational AI systems
- Developing a Retrieval-Augmented Generation (**RAG**)-based chatbot leveraging Nvidia's embedding models, Milvus vector database, and Gradio interface, to achieve at least 90% accuracy in real-time resume querying
- Implementing cloud-based infrastructure on **AWS** by leveraging services like S3 for candidate data storage, **RDS** for structured data management, and a vector database for embeddings, ensuring scalability and reliability

Machine Learning Engineer Aug. 2021 – Jul. 2023
PricewaterhouseCoopers *Bengaluru, India*

- Developed a scalable **FastAPI**-based backend to serve ML models for real-time financial document summarization and sentiment analysis using LangChain, attaining a ROUGE-2 score of 0.85
- Developed a centralized platform using Linux Server and PostgreSQL to unify data from Jira and Azure DevOps, reducing data hunting time by 50% and implemented CI/CD pipeline for automated deployment
- Containerized a multi-client product on Azure using **Docker** to streamline deployment, and monitor KPI, by automating REST API data ingestion, and orchestrating workflows with Apache Airflow DAGs
- Designed and implemented an efficient ETL workflow using Alteryx, streamlining data extraction, transformation, and loading processes, resulting in a 25% reduction in processing time and a 20% increase in data accuracy
- Collaborated on an end-to-end **MLOps** pipeline employing MLflow for model tracking and versioning to preprocess datasets with **10M+** data points and built time series models on Azure Databricks, accelerating development
- Outperformed baseline models by developing advanced models (XGBoost, LSTM) on **Azure Databricks**, reducing RMSE by 15% and MAE by 12%, resulting in a **20%** improvement in forecast accuracy using early stopping
- Orchestrated large-scale data workflows with **PySpark** and implemented production-level model retraining pipelines with MLflow, tripling retraining speeds and boosted real-time decision-making by 15%
- Delivered an AI-driven churn prediction system using BigQuery ML, XGBoost, and Google **Vertex AI**, elevating user retention rates and facilitating real-time predictions for a user base exceeding 1 billion

Projects

Brain MRI Classifier | *Link* Jan. 2025 – Feb. 2025

- Developed an end-to-end deep learning pipeline for multi-class brain MRI image classification, achieving an accuracy of **97%** using VGGNet and transfer learning along with MLOps tools like DVC and GitHub Actions
- Deployed the model on AWS using Docker and CI/CD pipelines, enabling scalable real-time image classification with seamless data access from AWS S3, processing over 7,000 MRI images efficiently.

Waymo 3D semantic segmentation | *Link* Aug. 2024 - Dec 2024

- Achieved a final mIoU of **0.70** and mAcc of **81%** on 1TB Waymo Open Dataset by optimizing the PointceptV3 semantic segmentation transformer for large-scale LiDAR data, enhancing performance across 23 class labels
- Improved mIoU by **22%** on ScanNetPP through fine-tuning, combining ScanNetPP training with Waymo dataset and utilized the high-performance computing units and managed GPU nodes for optimized performance