Venkatesh Shanmugam

Virginia US | svenkatesh.js@gmail.com | +1 (703) 216-2540 | GitHub
https://www.linkedin.com/in/svenkatesh-js/ | Portfolio

SUMMARY

Machine Learning Engineer with **4+ years of experience** delivering scalable AI/ML pipelines and deploying production-grade models on **AWS/GCP**. Skilled in **Python**, **TensorFlow**, **PyTorch**, and **MLOps** (**Vertex AI**, **Kubernetes**, **MLflow**, **Docker**). Proven record of achieving **78% predictive accuracy**, reducing pipeline latency to less than **200ms**, and optimizing **Spark-based ETL** (500M+ rows), saving costs by **40%**. Currently completing an **M.S. in Computer Science**, focused on applied ML, NLP, and responsible AI.

WORK EXPERIENCE

Senior AI/ML Engineer | ScriptChain Health | Washington, DC **May 2024 – Present**

Tech Stack: Python, PyTorch, TensorFlow, Vertex AI (GCP), Docker, Kubernetes (GKE), MLflow, Spark, SQL, A/B Testing

- Predicted 30-day hospital readmissions using research-based deep learning models with 78% accuracy, enabling earlier interventions across HIPAA-compliant hospitals.
- Cut deployment time from **4 hours to under 15 minutes** using Cloud Build, maintaining **P95 latency ≤200ms at 1K RPS.**
- Automatically promoted top performing models via **MLflow** A/B tests in Vertex AI, improving **PR AUC by +6pp** and cutting false positives by 8%.
- Rebuilt Spark ETL pipelines over 500M+ rows on AWS (S3, Glue, Athena), slashing preprocessing time by **99%** and GPU memory usage by **50%**, saving **$6K/year.**
- **Enabled** continuous retraining via MLflow + Cloud Logging with drift detection thresholds to maintain SLA reliability**.**
- **Partnered** with clinicians to define use cases that accelerated interventions and improved patient outcomes**.**

Data Consultant | George Washington University | Washington DC **September 2024 – Present**

Tech Stack: Pandas, NumPy, Scikit-learn, R, Jupyter Notebooks, Matplotlib, Seaborn, Excel, CSV data wrangling

- **Guided** 10+ research study in economics, biology, and NLP by mentoring students in regression analysis, hypothesis testing, and cross-validation using **R and Python**.
- **Empowered** 100+ students and faculty through 5 applied workshops on ML modeling, evaluation, and responsible AI practices.

Digital Transformation Developer | Tata Consultancy Services | India **April 2021 - August 2023**

Tech Stack: Data cleaning, Feature engineering, Process automation, Business process optimization, Cross-functional collaboration

- **Automated** workflows with custom scripts and tools for 14+ client teams, reducing manual effort and cutting costs by 40%.
- **Analyzed** client requirements and delivered data-driven process improvements, achieving **100% adoption** across all usecases.
- **Collaborated** cross-functionally with stakeholders to align automation solutions with business KPIs and operational goals**.**

EDUCATION

Master of Science in Computer Science (3.88 / 4.0), George Washington University **Aug 2023 - May 2025**

Bachelor of Technology in Computer Science (3.5 /4.0), SRM University **Aug 2016 - May 2020**

TECHNICAL SKILLS

Programming & Libraries: Python (pandas, NumPy, scikit-learn, TensorFlow, PyTorch), SQL, C++
Machine Learning: Predictive Modeling, Deep Learning (CNNs, Transformers), NLP (LLMs, Embeddings, RAG)
MLOps & Cloud: AWS (S3, Glue, Athena, EC2), GCP, Docker, MLflow, CI/CD, Git, Linux, REST APIs, Kubernetes
Data Engineering: Data Preprocessing, ETL pipelines, Big Data Tools (Spark), Distributed Training, Model Monitoring & Evaluation

PROJECTS

Intelligent Building Code QA (NLP Retrieval-Augmented Generation)

Tech Stack: Python, LangChain, OpenAI API, Pinecone, Streamlit, RAG, TensorFlow, Transformers, Sentence Transformer, LLMs

- **Built** a custom Q&A system for construction codes using a **Retrieval-Augmented Generation** pipeline. Integrated a vector database (Pinecone) for semantic search over building regulations and **GPT-4** for answer generation.
- **Improved** response efficiency for complex code queries by **98%**, significantly reducing the time architects spent on manual lookup. Deployed the solution as an interactive web app for demonstration.

AI-Text Discriminator

Tech Stack: Python, HuggingFace Datasets, scikit-learn, Pandas & NumPy, Matplotlib / Seaborn, TensorFlow

- Built an end-to-end NLP pipeline to classify human-written vs. LLM-generated text by **fine-tuning** transformer models **(BERT, GPT detectors)** on a custom dataset of 1.2 million text, achieving **97% accuracy** through robust training, hyperparameter tuning, and ensemble model stacking.
- Engineered features based on sentence embeddings, syntactic structure, and token distribution using **SentenceTransformer** and TensorFlow; evaluated model confidence thresholds for real-world deployment and integrated insights into a dashboard.