VARUN KUMAR KOTHAPALLI

Saint Louis, MO (Open to Relocate) | +1 (314) 556-4833 | Varun88645@gmail.com | LinkedIn | Portfolio

PROFESSIONAL SUMMARY

- AI / Machine Learning Engineer with 5+ years of experience designing, building, and operating production-ready machine learning systems across enterprise environments.
- Strong background in applied machine learning, including supervised and unsupervised modeling, feature engineering, model training, evaluation, and performance optimization using Python and industry-standard ML frameworks.
- Proven experience delivering Generative AI and NLP solutions leveraging Large Language Models (LLMs), Retrieval-Augmented Generation (RAG), and transformer-based architectures for document intelligence.
- Hands-on expertise in deploying, monitoring, and scaling ML models using MLOps practices, cloud platforms (AWS, Azure), containerization (Docker, Kubernetes), and RESTful APIs.
- Adept at collaborating with cross-functional teams to translate business requirements into reliable, scalable AI solutions with measurable operational impact.

TECHNICAL SKILLS

Artificial Intelligence & Machine Learning: Artificial Intelligence (AI), Machine Learning (ML), Deep Learning, Supervised & Unsupervised Learning, Feature Engineering, Model Training & Evaluation, Hyperparameter Optimization, Model Explainability, Bias Mitigation

Generative AI & Large Language Models: Generative AI, Large Language Models (LLMs), Transformer Architectures, Prompt Engineering, Fine-Tuning (LoRA, PEFT), Retrieval-Augmented Generation (RAG), Embeddings, Vector Search

Natural Language Processing (NLP): Text Classification, Named Entity Recognition (NER), Sentiment Analysis, Text Summarization, Text Generation, Semantic Search, Tokenization

Programming & ML Frameworks: Python, SQL, NumPy, Pandas, Scikit-learn, PyTorch, TensorFlow

MLOps, Deployment & Model Serving: MLOps, MLflow, Model Versioning, Model Deployment, Model Monitoring, Data Drift Detection, RESTful APIs, FastAPI, Flask, JSON APIs

Cloud Platforms & Containerization: AWS (SageMaker, EC2, S3, RDS), Azure Machine Learning, Docker, Kubernetes

Data Engineering for Machine Learning: ETL Pipelines, Feature Pipelines, Apache Spark, Data Warehousing

Databases: PostgreSQL, MySQL, Oracle, SQL Server

PROFESSIONAL EXPERIENCE

AI / Machine Learning Engineer **Mar 2025 - Present**
Equifax | Saint Louis, Missouri

- Designed Python- and SQL-based machine learning pipelines with Scikit-learn that streamlined feature preparation for large consumer credit datasets, cutting recurring model build cycles by around 30%.
- Strengthened credit risk and decisioning accuracy by training supervised and unsupervised ML models, producing 18-22% gains in accuracy and recall through disciplined cross-validation and tuning.
- Delivered real-time scoring capability by serving trained models through REST APIs built with FastAPI and Flask, enabling consistent prediction access across internal Equifax platforms.
- Stabilized production performance by introducing evaluation, monitoring, and drift-detection workflows, which reduced unexpected model degradation incidents by approximately 25%.
- Converted regulatory and analytics requirements into deployable AI decision logic, allowing model outputs to directly support KPIs tied to accuracy, compliance, and operational efficiency.
- Reduced manual document review effort by applying Generative AI and transformer-based NLP models for summarization and content extraction, achieving nearly 35% efficiency improvement.
- Improved release reliability by containerizing ML services with Docker and running inference workloads on AWS EC2, lowering environment-related deployment failures by about 40%.
- Sustained continuous model improvement through iterative development and review cycles supported by Azure Machine Learning experiment tracking and monitoring, enabling faster feedback integration and steady optimization of production ML systems.

Database Developer - Applied Machine Learning Systems **Sep 2021 - Jul 2023**
MedPlus Health Systems | Chennai, India

- Designed SQL-based ETL pipelines to consolidate pharmacy, prescription, and sales data, increasing the availability of ML-ready datasets for analytics and modeling by approximately 35%.
- Converted high-volume transactional records into stable ML feature tables through SQL-driven feature engineering, enabling consistent and reusable inputs for forecasting and analytical models.
- Reduced training data errors by around 30% by embedding validation rules and anomaly detection logic directly into ingestion and preprocessing workflows.
- Improved model experimentation speed by optimizing database schemas, indexing, and partitioning strategies, cutting feature retrieval latency by nearly 40%.
- Shortened data refresh cycles supporting model training by tuning SQL execution plans, allowing analytics and ML teams to iterate on models more frequently.
- Provided standardized dataset access by exposing curated healthcare data through JSON-based REST APIs, enabling seamless consumption by analytics and machine learning services.

- Protected sensitive healthcare and transactional information by enforcing data integrity checks and access controls, maintaining compliance across ML data pipelines.
- Aligned database design with evolving ML lifecycle and MLOps needs through close collaboration with data scientists and ML engineers, supporting scalable feature pipelines.

Software Engineer - Data Platforms **Mar 2019 - Aug 2021**
DXC Technologies India | India

- High-volume analytics workloads processed more efficiently once SQL Server queries, indexes, and stored procedures were optimized, increasing overall data throughput by around 30%.
- Reliability of datasets consumed by analytics and ML teams improved after introducing SQL-based preprocessing and aggregation logic, which reduced downstream data discrepancies by approximately 25%.
- Faster access to dashboards and model-preparation datasets followed the restructuring of database schemas and execution plans, cutting query response times by nearly 40%.
- Scalable compute and storage for enterprise reporting systems was enabled by supporting AWS EC2 and S3 deployments, improving availability of cloud-hosted analytics applications.
- Data platforms were better positioned for future AI and ML initiatives through active participation in architecture design discussions, ensuring warehouse structures aligned with evolving analytical needs.
- Unplanned outages became less frequent after performing root-cause analysis on data pipeline and reporting failures, reducing repeat incidents by about 20%.
- Clearer and more accurate business insights were delivered by converting client requirements into well-structured SQL data models, strengthening the quality of operational and performance reports.
- Knowledge continuity across delivery teams improved through detailed documentation of data flows, schemas, and processing logic, shortening onboarding time and reducing reliance on individual contributors.

PROJECTS

GENERATIVE AI-POWERED DOCUMENT INTELLIGENCE SYSTEM

- Implemented Generative AI and Large Language Models (LLMs) to automate document understanding and summarization, enabling accurate extraction of contextual insights from unstructured enterprise documents.
- Response relevance and factual grounding improved by integrating Retrieval-Augmented Generation (RAG) with embedding models and vector search, ensuring LLM outputs referenced indexed enterprise data.
- Inference reliability at scale was achieved by serving models through FastAPI, containerizing with Docker, deploying on AWS and registering models via MLflow with SageMaker-compatible artifacts, enabling reproducible and versioned deployments.

END-TO-END MACHINE LEARNING PIPELINE FOR PREDICTIVE ANALYTICS

- Structured datasets were prepared for modeling by orchestrating data ingestion, feature engineering, and preprocessing pipelines using Python, SQL, and Apache Spark, supporting repeatable ML experimentation.
- Predictive performance was improved through training supervised machine learning models with Scikit-learn and PyTorch, applying cross-validation and hyperparameter optimization techniques.
- Production stability and transparency were maintained by deploying models with Docker and Kubernetes, while implementing model evaluation, explainability, and monitoring aligned with MLOps practices.

EDUCATION

Master of Science in Information Technology **Aug 2023 - Dec 2024**
Webster University | Missouri, USA

Bachelor of Science in Computer Science **Jun 2015 - May 2018**
Loyola Academy Degree & PG College | Hyderabad, India

CERTIFICATIONS

- Machine Learning Specialization - **Coursera**
- Deep Learning Specialization - **Coursera**
- Generative AI with Large Language Models - **Coursera**
- Applied Machine Learning with Python - **Coursera**
- MLOps Fundamentals - **DataCamp**