POOJA BARALU UMESH

https://www.linkedin.com/in/pooja-baraluumesh

Bay Area, CA, USA | +16692491322

No Sponsorship needed | poojabumesh@gmail.com

SUMMARY

I'm a Machine Learning Engineer with **5+ years of experience** & Masters's in Data Science, working on products within various business domains, focusing on Distributed Systems, Cloud Platforms, Machine Learning, Data Science, and Data Mining. I possess expertise in designing, problem-solving, debugging, and analyzing requirements to achieve software performance and efficiency. My approach is result-oriented and hands-on, effectively managing resources and time constraints to deliver the best solutions.

SKILLS

Languages: Proficient in **Python** and SQL.
Machine Learning: Scikit-learn, TensorFlow, **PyTorch**, MLflow, Hugging Face Transformers
Distributed Systems: Apache Spark (SparkSQL, DataFrames), Hive, Presto, Airflow, Databricks, Snowflake
Cloud Platforms: Amazon Web Services (EC2, S3, Cloud Formation, Glue, Lambda, Stepfunctions, Sagemaker, ECS, EKS, ECR), Google Cloud Platform(Cloud storage, Big Query, MLFlow, VertexAI)
Databases: MongoDB, Pinecone (Vector DB), NoSQL
ML Concepts: Vector Embeddings, LangChain, **LLMs**, Prompt Engineering, NLP, Feature Engineering, Evaluation Frameworks, Time Series Forecasting, Clustering, Dimensionality Reduction (PCA), Content-Based Recommendation, RAG
Statistics & Analysis: ANOVA, Regression, GLM, Hypothesis Testing, Experimental Design
Visualization & Apps: PowerBi, Tableau**,** Matplotlib, Seaborn, Streamlit
APIs & Tools: SpaCy, NLTK, Cohere API

CERTIFICATIONS

AWS Certified Machine Learning - Associate, Amazon Web Services (AWS) Issued Oct 2025, **Badge**

PROFESSIONAL EXPERIENCE

Freelancer, Machine Learning August, 2025 - Present

- Led the development of AI-driven systems to optimize manufacturing processes by leveraging IoT, sensor, and ERP data, focused on building predictive intelligence into production workflows to enhance product consistency and operational efficiency.
  - Led research and prototype development of an AI-powered recommendation engine to suggest optimal process adjustments (temperature, time, ingredient ratios) for maintaining product parameters within target limits.
  - Designed and implemented data ingestion and ETL frameworks using MQTT, Snowflake, and Streamlit to automate collection and transformation of IoT, sensor, and ERP data into a unified data lake for model training and continuous optimization.

Drinks, San Jose, CA October, 2024 - July, 2025

Machine Learning Engineer

- I've designed and deployed scalable AI and data infrastructure solutions from conception to production, integrating LLMs, distributed data systems, and advanced MLOps workflows. My work spans RAG systems, real-time ML pipelines and distributed microservices across AWS and Kubernetes environments, bridging model innovation with production-grade reliability and cost efficiency.
  - Architected and deployed a scalable Retrieval-Augmented Generation (RAG) product-search platform on AWS SageMaker + ECS, integrating OpenAI LLMs, Pinecone vector DB, and Snowflake metadata store to enable context-aware, multi-tenant semantic retrieval across millions of SKUs.
  - Implemented an LLM-driven intent classification service using Transformers (PyTorch / vLLM) for semantic understanding of user queries; leveraged Spark on EMR for large-scale embedding generation and Kafka for real-time ingestion of catalog updates.
  - Enhanced ranking precision via a Hugging Face cross-encoder reranker and Cohere Reranker, applying cosine similarity + Levenshtein distance for fuzzy-match and feature-level attribute extraction using XGBoost and scikit-learn models.
  - Optimized inference pipelines through GPU-accelerated post-training tuning, caching strategies in AWS S3 and NoSQL (Cassandra / DynamoDB), and retrieval latency reduction of ~80 % using distributed compute orchestration in Kubeflow.
  - Developed an MLflow-based experiment-tracking and CI/CD framework integrated with snowflake and Presto, automating data ingestion, evaluation, and model promotion workflows through GitHub Actions + ECS deployments.

- Built a quantitative evaluation suite of 250 benchmark queries in Snowflake + S3, enabling continuous performance monitoring, regression detection, and automated model-quality reporting via FastAPI microservices.
- Partnered directly with the CTO and Co-founder to present architecture design reviews and production metrics, driving organization-wide adoption of the RAG platform and seamless API integration into the company's e-commerce stack

---

Harmony Food Pvt. Ltd., India March, 2022 - January, 2024

Software Engineer

- I designed and built data pipelines and analytics systems that streamlined manufacturing operations, improved production visibility, and enhanced decision-making across the supply chain. My work bridged raw production data, quality metrics, and business intelligence to enable real-time performance tracking and process optimization.
  - Architected and deployed end-to-end ETL pipelines on AWS EMR and Apache Spark, orchestrated via Airflow DAGs, to process petabyte-scale production and quality-control datasets, reducing manual reporting time by 60 % and improving system reliability under high-volume workloads.
  - Designed distributed data ingestion workflows using Apache NiFi, Presto, and S3-based staging zones to unify supplier, logistics, and IoT feeds into a central Snowflake warehouse, ensuring sub-minute data latency and strong schema consistency.
  - Developed high-performance Python modules and FastAPI microservices for metadata tracking, data validation, and anomaly detection; automated quality-alert pipelines that cut issue-response time by 70 % across teams.
  - Integrated supplier and logistics APIs into the central warehouse database, increasing end-to-end supply chain visibility and data reliability.
  - Built and maintained interactive analytics dashboards in Apache Superset and Power BI, integrating with Presto and Hive tables for near-real-time visibility into inventory health, yield metrics, and supplier SLAs used by operations leadership.
  - Implemented CI/CD workflows on GitHub Actions and Kubernetes, automating Spark job deployments, schema migrations, and regression tests to maintain consistent code quality and reproducibility across environments.
  - Optimized data-processing performance through partition pruning, adaptive Spark execution, and S3 I/O tuning—achieving up to 40 % cost savings on EMR compute and improving query response times by 3×.
  - Developed a scalable business-analytics layer leveraging Presto, HiveQL, and Parquet to support ad-hoc reporting and KPI computation for finance and operations teams.
  - Mentored and onboarded junior data engineers, leading deep-dive sessions on Airflow pipelines, distributed systems, and AWS data-lake design, while delivering technical talks on data quality and performance best practices. Collaborated cross-functionally with analysts, ML engineers, and PMs to design scalable data models and workflows, fostering a data-driven culture company-wide.

Anheuser-Busch, India October, 2020 - January, 2022

Assistant Manager

- I led data-driven manufacturing optimization initiatives, transforming raw sensor data into actionable insights to improve efficiency, sustainability, and operational visibility. My work combined analytics, automation, and time-series modeling to drive measurable process improvements across production lines.
  - Led a cross-functional analytics project that reduced extract loss by 9% and water usage by 5% using **time-series anomaly** detection on brewhouse sensor data.
  - Developed and maintained SQL-based data pipelines to monitor Overall Equipment Effectiveness (OEE), enabling near real-time tracking and contributing to a 25% increase in productivity.
  - Automated **report generation and KPI tracking** by integrating data from **Manufacturing Execution Systems (MES)** into interactive **dashboard visualizations** for leadership review.

IFFCO, UAE October, 2017 - August, 2018

Data Analyst

- Analyzed production and quality datasets using SQL and Excel to identify process optimizations, increasing production parameters by 20% through additive ratio tuning.
- Built automated GMP compliance reports in Power BI to track audit metrics and streamline documentation workflows, improving data accuracy and audit readiness.

EDUCATION

University of San Francisco, CA July-2025
Masters in Data Science

PROJECTS

Developed A(I)YE Chef, an end-to-end AI-powered culinary assistant Github

- Fine-tuned YOLOv8 on 24k+ images for 120-class ingredient detection (>95% accuracy).
- Integrated Vertex AI Gemini LLM to generate JSON-structured personalized recipes.
- Exported and deployed the YOLOv8 PyTorch model via FastAPI in a Docker container, leveraging GCP Cloud Run for scalable, serverless inference.
- Implemented MLflow for artifact logging and model registration, resolving conflicts with YOLO's auto-logging to centralize metric tracking and optimize compute costs through transfer learning and serverless deployment.

Tweet Popularity Predictor – End-to-end ML pipeline for social media analytics Github, Blog

- Multi-task pipeline for emotion classification (DistilBERT), hashtag generation (GPT-2), and popularity scoring (linear regression).
- Designed retrainable Python package with CLI & API; integrated with Snowflake for storage & dashboards.
- Optimized inference speed via batch processing, reducing runtime for large datasets; added unit test scaffolding and roadmap for FastAPI microservice deployment.

Fine-Tuning Stable Diffusion 2.1 for Domain-Focused Image Generation Blog

- Fine-tuned Stable Diffusion 2.1 on curated ArtBench-10 dataset, enabling stylistically coherent image generation aligned with domain-specific artistic prompts and style requirements.
- Designed robust preprocessing pipeline including deduplication via perceptual hashing, CLIP normalization, and UTF-8 text cleaning to ensure high-quality training data.
- Implemented memory-efficient training using WebDataset with 46 sharded tar files and LoRA-based PEFT on A100/4090 GPUs, optimizing resource utilization and training time.
- Evaluated model performance using CLIP similarity scores and human assessment, achieving improved prompt adherence and visual fidelity in generated artwork