PRATEESH PATLOLLA

Data Scientist / ML Engineer

(650) 474-5593 •
patlollaprateesh@gmail.com •
prateeshreddy.github.io •

SUMMARY

Machine Learning Engineer with 4+ years of experience delivering GenAI, NLP, Forecasting, and Optimization solutions at scale. Built cloud-native ML systems using AWS, Azure, Spark, and MLOps tools from experimentation to deployment. Delivered high-impact products including LLM-powered RAG assistants (LangChain, FAISS, LangGraph, OpenAgents, Bedrock), Gurobi-based production optimizers, and real-time APIs for cross-functional teams. Skilled in A/B testing, Causal inference, and Stakeholder-facing experimentation, with a track record of driving measurable ROI across enterprise environments.

SKILLS

- **Languages**: Python, SQL, R, Java, JavaScript
- **Big Data & Pipelines**: Apache Spark, PySpark, Hadoop, Hive, Parquet, AWS Kinesis, Kinesis Firehose, Glue
- **Cloud & MLOps**: AWS (SageMaker, Bedrock, Redshift, Lambda, CloudWatch), Azure (incl. Azure OpenAI), Docker, Kubernetes, Airflow, Jenkins, FastAPI, GitHub Actions, VPC Networking
- **GenAI & LLMs**: LangChain, LangGraph, OpenAgents, Guardrails AI, Llama 2, GPT-4, Prompt Engineering, Vector DBs (FAISS, Pinecone, Chroma), LLMOps tools (PromptLayer, TruLens)
- **Databases & Storage**: Redis (ElastiCache), DynamoDB, PostgreSQL, Oracle, Redshift, MongoDB
- **Analytics & BI**: Tableau, Power BI, AWS QuickSight
- **Modeling & Optimization**: Scikit-learn, XGBoost, SARIMAX, PyTorch, SHAP, Gurobi, A/B Testing, Causal Inference
- **Monitoring**: CloudWatch, Datadog

PROFESSIONAL EXPERIENCE

TOYOTA NORTH AMERICA – Plano, TX
Senior Data Scientist *February 2023 – Present*

- Led the design and POC of AskToyota, a GenAI chatbot using LangChain, FAISS, LangGraph, and LLaMA 2, deployed in secure VPC environments with IAM-based access control and TLS. Reduced stakeholder turnaround time by 30% by surfacing forecasts and KPIs on demand.
- Spearheaded Accessory Recommendation engine generating $150M+ annual revenue using XGBoost, dynamic segmentation, and incentive-aware retraining, integrated A/B testing and causal inference for robust evaluation.
- Built Toyota's Annual Production Optimizer (APO) using Gurobi to solve multi-objective LP problems (profit, volume, GHG score) across 14 global factories, optimizing $700M+ in planning annually.
- Developed Smart Forecasting engine with PySpark and SARIMAX to predict dealer-level sales; enabled 3+ hours/month time savings per dealer and built self-serve analytics dashboards via QuickSight.
- Productionized ML pipelines on SageMaker, integrating model registry, drift detection with auto-retraining, Airflow orchestration, and CloudWatch/Datadog monitoring for real-time alerts.
- Deployed real-time ETA API for dealer websites using ElastiCache Redis with SHA-1 for online inference; batch-trained models using PySpark and Parquet on S3. Packaged with Docker and deployed via Kubernetes.
- Blended causal inference and A/B testing frameworks with GenAI tools (AWS Bedrock, LangChain agents, Guardrails AI, Pinecone, FAISS) to auto-explain accessory model KPIs, increasing dealer conversions by 20% and boosting stakeholder trust.

AMAZON ALEXA AI – Santa Barbara, CA
Data Scientist Intern *May 2022 – August 2022*

- Retrained and deployed Alexa's question categorizer using advanced NLP (multi-class classification, intent detection), improving query routing across 100+ categories via SageMaker.
- Used resampling and confidence intervals to evaluate and improve model confidence on rare query intents, increasing trust and accuracy for low-frequency categories in Alexa.
- Improved the generated success rate model's alignment with human-labeled success rate, reducing gap from ~20% to under 14%, enabling reduced annotation spend.
- Built reporting pipelines that pushed KPIs into Redshift and Quicksight to visualize key health metrics and rolling forecast 3-month trends to support weekly reviews with product stakeholders.

CYIENT – Hyderabad, India
Data Scientist *August 2019 – December 2020*

- Extracted road sign boards through object detection from terrestrial imagery to minimize manual efforts of data annotation for North American-based clients.
- Achieved a hit rate of 92%, resulting in a saving of 12 FTEs.
- Designed and optimized AWS Glue ETL jobs to streamline data ingestion into a data warehouse, leading to significant improvements in data quality and processing speed.

EDUCATION

INDIANA UNIVERSITY BLOOMINGTON **Bloomington, IN, USA**
Masters of Science in Data Science

GITAM UNIVERSITY **Hyderabad, India**
Bachelor's in Computer Science

PROJECTS

Prateesh's GenAI Resume Assistant *link: prateeshreddy.github.io*

- Developed a scalable Retrieval-Augmented Generation (RAG) pipeline using LangChain, FAISS, and OpenAI's embeddings, enabling context-aware, accurate Q&A about my professional background.
- Engineered advanced prompt templates and conversational memory, delivering enterprise-grade efficiency with 98% retrieval accuracy, 120 ms latency (p99), and inference costs below $0.0005 per query.
- Deployed serverlessly on Cloudflare Workers with robust CI/CD via GitHub Actions, featuring real-time observability and automated logging for iterative improvements.

Cognitive Search Engine by NLP

- Provided Cognitive search capability for Eli Lilly and Company to search against databases like FDA and EMA via natural language questions and return relevant results to help with accelerating regulatory submissions for Eli Lilly as an Intern in the Summer of 2021 using Hugging Face transformers models.

Research Scientist at Institutional Analytics, Indiana University

- Been part of the guest lecture for ILSZ 637 - Information Visualization by Noriko Hara in Spring 2022, a graduate-level course talking about data storytelling and data visualization.
- Built data pipelines with Tableau reports to analyze student performance, retention, and graduation trends, influencing key academic decisions.