Prateesh Patlolla

Data Scientist / ML Engineer

patlollaprateesh@gmail.com | +1 (650) 474 5593

SUMMARY

- Experienced Data Scientist and Machine Learning Engineer with 3+ years building real-time ML and optimization solutions in supply chain, production planning, and forecasting environments. Skilled in managing full ML lifecycle from experimentation to production deployment at scale using AWS, Spark, and MLOps practices.
- Proven expertise in targeting strategies, demand forecasting, statistical modeling, optimization, and building GenAI applications within large-scale enterprise environments. Adept at developing scalable ML pipelines, deploying LLM-driven solutions using LangChain and FAISS, applying explainable AI techniques, and optimizing data workflows across Redshift, PySpark, and big data ecosystems.

SKILLS

Languages: Python, SQL, R, Java, JavaScript, SAS
Big Data & Distributed Computing: Apache Spark, Hadoop, PySpark, Hive
Cloud & MLOps: AWS (SageMaker, Glue, Redshift), Azure, Docker (model deployment), FastAPI (API serving), CI/CD with Airflow, Jenkins, GitHub Actions
Databases & Data Warehousing: Oracle DB, Redshift, PostgreSQL, MySQL, SQL Server, MongoDB
Visualization & BI Tools: Tableau, Power BI, AWS QuickSight
Machine Learning & Modeling: Scikit-learn, XGBoost, PyTorch, SHAP, A/B Testing, Causal Inference
Operating Systems: Windows, Linux

EXPERIENCE

Toyota, NC | Feb 2023 – Present | Senior Data Scientist

- Spearheaded a $150M annual optimization initiative, leading end-to-end design, modeling, and deployment of predictive analytics pipelines to enhance Toyota's supply chain and production planning.
- Led predictive scheduling and production planning initiatives, developing scalable ML workflows that integrated supply chain constraints and production targets, improving monthly manufacturing schedules, reducing overstock, and optimizing plant-level throughput.
- Developed and deployed an internal GenAI assistant ("AskToyota") using LLaMA2, FAISS, and LangChain to streamline access to forecasting and scheduling documentation, improving operational decision-making speed and cross-team productivity.
- Built scalable ML pipelines using AWS SageMaker, Airflow, and CloudWatch, enabling automation and real-time monitoring while reducing operational overhead.
- Designed and implemented large-scale production planning solutions using constrained linear programming with Gurobi, improving plant-level efficiency and cost savings.
- Applied causal inference and A/B testing frameworks to evaluate ML impact and drive KPI development, resulting in a 20% lift in accessory conversions.
- Engineered big data pipelines using PySpark and Parquet, and wrote optimized SQL for Oracle DB and Redshift, improving query performance and data accessibility by over 30%.

Alexa AI Amazon, Santa Barbara, CA | May 2022 - Aug 2022 | Data Scientist Intern

- Improved the Alexa categorizer model by utilizing AWS Glue for data extraction and feature engineering, which enhanced prediction accuracy for low-frequency queries.
- Created a model of stratified sampling techniques to reduce the Confidence Interval, providing actionable insights into Alexa's performance and improving user satisfaction.
- Built a reporting pipeline and dashboards using AWS QuickSight for historical analysis of Alexa's success rate, which was adopted as a primary reporting tool in Alexa's weekly business webinars.

Cyient Ltd, India | Aug 2019 - Dec 2020 | Data Scientist

- 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 ETL processes to streamline data ingestion into a data warehouse, leading to significant improvements in data quality and processing speed.
- Developed comprehensive Power BI and Tableau dashboards for North American clients, leading to operational savings by providing critical insights into data patterns and business metrics.

EDUCATION

Masters of Science in Data Science | Indiana University Bloomington, Indiana, USA
Bachelors in Computer Science | GITAM University Hyderabad, India

PROJECTS

Prateesh's GenAI Resume Assistant:

- 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 huggingface 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.