SURAJ JAGTAP

Data Scientist | AI/ML Engineer

Mumbai, India | +91-7028143833 | jagtapsuraj636@gmail.com | LinkedIn | GitHub

---

1 Profile

- Highly accomplished Data Scientist and AI/ML Engineer with over 7 years of experience delivering enterprise-grade AI systems and MLOps pipelines across AWS and GCP environments.
- Specialized in LLMOps, foundation model tuning, and orchestrating multi-modal AI solutions, achieving 30% cost reduction and 40% faster delivery.
- Proficient in aligning AI initiatives with business metrics, ensuring ethical, transparent, and scalable implementations.
- Expert in predictive modeling, generative AI, large language models (LLMs), and explainable AI, with a focus on business impact and scalability.
- Experienced in managing end-to-end MLOps pipelines, CI/CD workflows, and cross-functional team collaboration.
- Strong knowledge of cloud platforms (AWS, GCP), containerization (Kubernetes, Docker), and observability tools (Prometheus, Grafana).
- Excellent communication, stakeholder management, and project delivery skills in agile and hybrid environments.

2 Technical Skills

Category	**Skills**
Programming	Python, SQL, PySpark, R
ML Frameworks	TensorFlow, PyTorch, Scikit-learn, Hugging Face, Keras, LLaMA, Mistral, LoRA
MLOps & Cloud	AWS (SageMaker, Lambda, S3, CloudWatch), GCP (AI Platform, Cloud Run), Kubernetes, MLflow, Terraform, Prometheus, Grafana
Data Tools	Pandas, NumPy, Spark, MongoDB, PostgreSQL, Power BI, Tableau, Looker
LLM Stack	GPT, BERT, LangChain, FAISS, Pinecone, Weaviate, Prompt Engineering, RAG
AI Best Practices	RLHF, SHAP, AutoML (H2O.ai), Explainable AI (XAI), Drift Detection, Model Monitoring

3 Certificates

- AWS Certified Machine Learning – Specialty | Amazon | 2024
- GCP Professional Machine Learning Engineer | Google | 2024
- Data Science Professional Certificate | IBM | 2024
- TensorFlow Developer Certificate | Google | 2023
- Generative AI with LLMs | Coursera (DeepLearning.AI) | In Progress (2025)

4 Professional Experience

Duration	Organization	Designation
July 2022 – Jan 2025	Modi Motors, Mumbai	Data Scientist
Jan 2022 – July 2022	Shaw Motors, Mumbai	Data Scientist
Jan 2019 – Jan 2022	NBS International Ltd, Mumbai	Data Scientist
Sep 2017 – Dec 2018	Salasar Autocrafts, Mumbai	Data Scientist
Apr 2015 – June 2017	Unitech Automobiles Ltd, Mumbai	Data Analyst

5 Project Experience – Details

Project 1
Project Name: Enterprise LLM Chatbot with RAG Architecture
Client: Modi Motors
Role: Data Scientist
Duration: July 2022 – Jan 2025
Environment: GPT-3.5, FAISS, Pinecone, LangChain, AWS, GCP
Team Size: 10
Description: Developed a scalable LLM-based chatbot using GPT-3.5 with RAG architecture and Pinecone for vector search, improving customer query accuracy by 35% and scaling to 1,000+ concurrent sessions. Integrated outputs with Power BI for real-time business insights, reducing extraction time by 50%. Applied RLHF, LoRA, and Quantization to reduce hallucinations by 15% and ensure >95% model reliability using AWS CloudWatch and GCP Logging.
Responsibilities:

- Led end-to-end development of the chatbot, from requirement gathering to deployment.
- Designed and implemented RAG architecture with Pinecone and FAISS for efficient retrieval.
- Fine-tuned LLaMA models using RLHF and LoRA to optimize performance and reduce latency.
- Integrated LLM outputs with Power BI dashboards for actionable business insights.
- Set up observability pipelines with AWS CloudWatch and GCP Logging for model monitoring.
- Collaborated with stakeholders for iterative feedback and agile delivery.

Project 2
Project Name: Vehicle Insurance Prediction Pipeline (MLOps)
Client: NBS International Ltd
Role: Data Scientist
Duration: Jan 2019 – Jan 2022
Environment: XGBoost, H2O.ai, FastAPI, GitHub Actions, AWS, GCP
Team Size: 8
Description: Built an end-to-end MLOps pipeline for vehicle insurance prediction, achieving a 20% improvement in model F1 score (0.89) using XGBoost and Optuna. Reduced model drift by 40% with real-time detection and alerting via CloudWatch, Prometheus, and Slack. Automated CI/CD workflows with GitHub Actions, cutting deployment lead time by 35%. Integrated MongoDB for data ingestion, YAML for validation, and SMOTE for data transformation.
Responsibilities:

- Gathered and validated requirements for the insurance prediction pipeline.
- Designed and implemented a modular MLOps pipeline with FastAPI microservices.
- Optimized XGBoost models using Optuna for hyperparameter tuning.
- Set up drift detection and automated alerts with CloudWatch and Prometheus.
- Managed CI/CD workflows via GitHub Actions for seamless deployments.
- Conducted sprint planning, stakeholder demos, and user acceptance testing (UAT).

Project 3
Project Name: AI-Powered Vehicle Valuation System
Client: Salasar Autocrafts
Role: Data Scientist
Duration: Sep 2017 – Dec 2018
Environment: Python, Pandas, Scikit-learn, Flask, Power BI
Team Size: 6
Description: Developed a web-based vehicle resale prediction system using Python and scikit-learn, improving pricing accuracy by 23% (R^2=0.87). Enhanced dealer engagement by 17% with explainable models using XGBoost and SHAP. Built a Flask-Power BI dashboard with 98.4% uptime for real-time insights, enabling faster and more accurate resale estimations.
Responsibilities:

- Gathered requirements and designed the valuation system's architecture.
- Developed predictive models using scikit-learn and XGBoost with SHAP for explainability.
- Built and deployed a Flask-based web application integrated with Power BI dashboards.
- Automated feature engineering and ensured scalability across dealer networks.
- Facilitated client feedback sessions and conducted UAT for deployment readiness.
- Maintained project documentation and ensured alignment with business goals.

6 Education

Qualification	University/Board	Duration
B.E. – Mechanical Engineering	Savitribai Phule Pune University	2016 – 2020
MBA – Business Analytics	Shree Shivaji Maratha Society's Institute of Management Studies, Pune	2021 – 2023