Sumit Shinde

+91-848-293-5657 | shindesb1615@gmail.com | linkedin.com/in/sumit-shinde | github.com/shindesb1615

Education

Vishwakarma University — Pune, India
B.Tech in Computer Engineering; CGPA: 7.64 — *2024 – Present*

Shree Ramchandra College of Engineering — Pune, India
Diploma in Computer Engineering; Percentage: 75.66% — *Passed 2024*

Experience

Data Science Intern — Apr 2024 – Jun 2024
Shorat Innovations — *Pune, India*

- Performed in-depth exploratory data analysis on 100k+ structured records using Python (Pandas, NumPy), uncovering key data patterns that informed feature selection and model development.
- Designed interactive dashboards using Plotly and Seaborn to visualize temporal and categorical trends, reducing manual reporting time by 60% and aiding data-driven decision-making.
- Engineered lightweight web scraping pipelines using BeautifulSoup and Requests to automate data collection from 50+ webpages.

Projects

Multimodal E-Commerce Product Intelligence | *CLIP, BLIP-2, LLaVA, Computer Vision, Responsible AI*

- Designed and benchmarked a multimodal AI pipeline (image → product category and description) evaluating transformer-based architectures including CLIP, BLIP-2, and LLaVA, and recommended a hybrid pipeline for production deployment.
- Incorporated Responsible AI practices, including bias and fairness checks across model outputs, to support equitable category and description generation.
- Authored a technical report detailing architecture tradeoffs, accuracy, and latency considerations to guide model selection for production use.

YouTube Analytics & Chat Assistant | *Chrome Extension, AI/ML, MLOps, AWS*

- Developed a sentiment analysis model using LightGBM (87% accuracy), integrated T5-small for video summarization, and constructed a RAG chatbot using FAISS for semantic search.
- Prepared a CI/CD pipeline with GitHub Actions, MLflow for experiment tracking, DVC for data versioning, and automated testing achieving 95%+ test coverage using PyTest.
- Deployed a Flask-based backend on AWS leveraging Docker, ECR, CodeDeploy, Auto Scaling Groups, and Load Balancers to ensure scalability and reliability; implemented asynchronous request handling to support concurrent comment-analysis requests without blocking.
- Created a Chrome Extension as the frontend interface, integrated with 8+ Flask REST API endpoints delivering real-time comment analysis, sentiment trends, and AI-driven insights.
- Implemented interactive word clouds, sentiment graphs, comment summarization via LLM, contextual chatbot memory, and real-time article enrichment using the Tavily API.

NYC Taxi Demand Prediction | *End-to-End ML Pipeline, MLOps, Cloud Deployment*

- Built a demand forecasting system for NYC Yellow Taxis using 3M+ trip records, predicting 15-minute interval demand across 30 zones.
- Engineered time-series features such as EWMA (80% improvement in accuracy) and clustered regions using Mini-Batch KMeans.
- Trained a Linear Regression model (92.1% accuracy) with real-time zone-level demand visualizations in a Streamlit dashboard.
- Containerized with Docker and deployed on AWS (EC2, S3, ECR, CodeDeploy) using GitHub Actions for CI/CD, with asynchronous/batch processing to handle large-scale data ingestion and feature computation; used PyTest for test automation.
- Tracked model iterations with MLflow and versioned datasets (5GB+) using DVC to ensure reproducibility and safe rollback during deployment cycles.

Technical Skills

Programming & ML: Python (Pandas, NumPy, scikit-learn, PyTorch, TensorFlow, Matplotlib, Seaborn), SQL, Java, JavaScript, Bash, Git/GitHub, Statistical Analysis, Supervised/Unsupervised Learning, Deep Learning, NLP, Ensemble Methods (LightGBM, XGBoost)
Generative AI & LLMOps: RAG, LangChain, LangGraph, LlamaIndex, Hugging Face, Prompt Engineering, Agentic AI, Context Engineering, Fine-Tuning (PEFT, RLHF), Multimodal Pipelines (CLIP, BLIP-2, LLaVA), Vector DBs (FAISS, AstraDB, Pinecone), MLflow, DVC, vLLM, W&B
Responsible AI: Bias Detection, Fairness Evaluation, Model Governance
Data Engineering & Cloud: ETL Pipelines, Apache Airflow, PySpark, DBT, Snowflake, Kafka, RabbitMQ, AWS (EC2, S3, ECR, Lambda, Auto Scaling), GCP, Azure, Docker, Kubernetes, PostgreSQL, MySQL
DevOps & Production: CI/CD (GitHub Actions, Jenkins, CircleCI, ArgoCD), REST APIs (Flask, FastAPI), Asynchronous & Distributed Processing, PyTest, Prometheus, Grafana, ELK Stack, SonarQube, Seldon Core
Automation & Low-Code: n8n, Zapier, Make
Security & Auth: OAuth, SAML, SSO

Certifications

IBM Full Stack Developer – Coursera (2025)
Developing AI Applications with Python and Flask – Coursera (2026)