# Shushant Kamatar

8088348874 | shushantkamatar11@gmail.com | LinkedIn | Github

## Experience

**Machine Learning Engineer (Contract)** Nov 2025 - Present

*Forin Agency* *Remote*

- Architected a real-time Interview Intelligence System using FastAPI and WebSockets, achieving 800ms latency by parallelizing audio transcription (Deepgram) and LLM inference (Llama 3).
- Engineered a context-aware RAG pipeline with Supabase (pgvector) to dynamically retrieve candidate resume data and historical "winning questions" during live calls.
- Developed a secure Chrome Extension integration using the 'Offscreen' API pattern to capture high-fidelity audio streams from Google Meet while bypassing browser sandbox restrictions.

**Software Developer** May 2024 – Sept 2025

*Lenovo* *Bengaluru*

- Engineered a full-stack sentiment analysis dashboard for Lenovo products, integrating a Python backend with Flask for a dynamic frontend and live NLP results.
- Automated critical file merging workflow with Python and pandas, reducing runtime by 99.9% and significantly boosting team productivity.
- Spearheaded rollouts and project implementations across Western Europe, ensuring successful deployment and user adoption in 15 countries.

## Projects

**Distributed Multi-Model Inference Verification Gateway**

- Architected a high-concurrency inference gateway aggregating multiple LLM providers (Groq, Gemini), reducing end-to-end latency by 40% via asynchronous parallelization and ensuring 99.9% service availability through automated provider failover strategies.
- Pioneered a "self-healing" code execution engine that autonomously detects, debugs, and patches runtime errors in a sandboxed environment, increasing successful code generation rates from 65% to 92% without human intervention.
- Engineered a hybrid semantic caching layer using Redis and vector embeddings that reduced external API costs by 45% and improved P99 latency for recurring queries by 600ms, optimizing performance for resource-constrained deployments.

**ProductPulse AI: End-to-End Generative AI Analysis Platform**

- Deployed a Dockerized FastAPI (Uvicorn) microservice to Render, proving a production-ready MLOps workflow by solving cloud credential and worker boot failures.
- Engineered a LangChain (LCEL) pipeline using the Google Gemini API to analyze 100+ text reviews, automatically extracting 3 key themes and actionable business insights.
- Built and deployed an interactive Streamlit dashboard that consumes the live API, providing instant AI-driven analysis to reduce research time from hours to seconds

## Technical Skills

**Languages**: Python, SQL , JavaScript
**Machine Learning & GenAI** : LLMs (Llama 3, Gemini), CNN, Scikit-Learn
**MLOps & Cloud**: Docker, Kubernetes (K8s), GitHub Actions (CI/CD), Prometheus, Grafana, MLflow, DVC, AWS
**System Design & Backend**: Asynchronous Programming (AsyncIO), REST APIs (FastAPI), Redis (Caching)
**Data Science & Analytics**: Predictive Modeling, A/B Testing, Statistical Analysis, EDA, Matplotlib, Seaborn

## Education

**Bosscoder Academy**

*Datascience and Machine Learning*

**RV College of Engineering**

*BE in Aerospace Engineering* *June 2019 – Aug 2023*

## Certifications

**Nasscom Certified Data Scientist**