Shreyansh Kumar AI/ML Engineer

Gmail | (617) 259-0101 | Las Vegas, NV | LinkedIn | GitHub

Summary

AI/ML Engineer with 2.5+ years of strong expertise in machine learning, deep learning, and data analytics. Skilled in building scalable models, data pipelines, and cloud-based deployments. Proficient in Python and advanced ML frameworks, with experience across the full model lifecycle. Adept at cross-functional collaboration, automation, and delivering data-driven solutions to support business decision-making and improve efficiency.

Education

Master of Science in Applied Data Analytics 09/2023 – 01/2025
Boston University, Boston, MA, USA
Bachelor of Technology (B.Tech), Computer Science 08/2019 – 03/2023
Bennett University Greater Noida, Uttar Pradesh, India

Technical Skills

- Programming Languages: Python, SQL, C++, R
- ML & AI Systems: Retrieval-Augmented Generation (RAG), Large Language Models (LLMs), NLP, Computer Vision,Classification, Regression, Clustering, Gradient Boosting
- Frameworks & Libraries: PyTorch, TensorFlow, Scikit-learn, Keras, XGBoost, Hugging Face, LangChain, CrewAI
- Data & Distributed Systems: Apache Spark, Google BigQuery, Vector Databases (Weaviate, Chroma), Big DataCloud & Deployment: GCP, AWS, Azure, Streamlit, REST-based ML services
- Analytics & Experimentation: A/B Testing, KPI Design, Experimentation, Model Evaluation

PROFESSIONAL EXPERIENCE

Droisys | AI/ML Engineer | Las Vegas, NV, USA 02/2026 – Present

- Hardened observability across a multi-module Spring Boot codebase by replacing ad hoc console output with SLF4J structured logging, aligning application diagnostics with Logback file and console pipelines to support production troubleshooting and audit-friendly traces
- Contributed to the Table Guard "Web API" (GUIManager) exposing REST resources for alerts, players, gaming, and configuration, integrating JWT-secured routes with shared services (AlertService, PlayerService, GamingService, ConfigurationService) over PostgreSQL for operational casino surveillance workflows
- Supported data-intensive pipelines by working in an ETL + rule-engine architecture that ingests and transforms table game data and evaluates alert rules (e.g. Blackjack/Baccarat contexts), connecting batch processing and domain logic used to surface risk signals to operators

Chicago Education Advocacy Cooperative (CHiEAC) | Lead Data Analyst | Remote, USA 03/2025 – 02/2026

- Designed an AI-powered, multi-agent tutoring platform, replacing static study material with adaptive explanations, quizzes, and revision workflows, resulting in 30% improvement in learner outcomes and 35% higher student engagement
- Architected a scalable retrieval and reasoning pipeline using recursive content chunking, optimized embeddings, and metadata-aware vector search, improving answer relevance and reducing topic-level confusion by 42%
- Built a feedback-driven prompt tuning and evaluation loop, enabling the tutor to continuously adapt to repeated learner queries while reducing teacher intervention and aligning improvements with measurable academic outcomes
- Developed an end-to-end intelligent document processing system combining computer vision enhancement, OCR, and hybrid rule-based + LLM extraction, achieving 100% processing success, 98% extraction accuracy gains, and 90% reduction in API costs through intelligent method selection

Nexdigm | Data Scientist | Gurugram, India 08/2022 – 06/2023

- Led development of a Python-based time series tool using ARIMA, SARIMA, SARIMAX, VARMAX, and RNN models, reducing exploratory forecasting time by 70% through efficient automation and streamlined analysis workflows
- Automated model selection and hyperparameter tuning, running 50+ iterations per dataset, which significantly reduced project timelines from 3 weeks to just 4 days, accelerating delivery cycles for multiple time-sensitive client projects
- Enhanced model accuracy by 12% using iterative tuning, advanced feature engineering, and performance optimization, enabling more reliable predictions and data-driven decision-making for stakeholders across various industries and use-case domains
- Boosted client engagement by 15% through timely delivery of predictive insights, leveraging automated time series forecasting outputs to support strategic planning, drive value, and strengthen ongoing relationships with enterprise customers

TransOrg Analytics | Data Analyst Intern | Gurugram, India 02/2022 – 06/2022

- Predicted sales using advanced time-series models (ARIMA, Prophet) to drive data-informed inventory decisions, resulting in a 12% reduction in overstock and a 9% decrease in stockouts, improving operational efficiency and product availability
- Improved product discovery and user engagement by implementing semantic search powered by embedding models, leading to a 30% increase in average session duration and a higher interaction rate across key product categories
- Boosted repeat purchase rates by 18% by developing dynamic customer segments using clustering algorithms (K-Means, DBSCAN) and deploying personalized recommendations and email campaigns tailored to user behaviour and preference

PROJECTS

Big Data Based Dining Recommendation System | Vector Databases, NLP, RAG, LangChain, Semantic Search

- Built a large-scale retrieval-augmented recommendation system over 8GB+ Yelp dataset, delivering context-aware dining
- recommendations with 95% response accuracy through semantic retrieval and LLM-based generation
- Designed a scalable ML inference pipeline using BigQuery-backed data processing and vector search, supporting real-
- time recommendations across 10K+ businesses with low-latency querying

Health Risk Classification Pipeline | Feature Engineering, Dimensionality Reduction, Hyperparameter Tuning, ML Pipeline

- Built an end-to-end classification pipeline with feature engineering and tuning, reaching 70.93% weighted accuracy and 79.51% true positive rate, and raised predictive performance ~18% using SMOTE for class imbalance and Boruta for feature selection.
- Cut model training time ~30% with dimensionality reduction (Lasso, ANOVA F-tests) while holding accuracy steady, improving iteration speed on noisy, imbalanced data.

Certificates

Microsoft AZ-900 | Microsoft AI 900 | Oracle Cloud Infrastructure 2021 (Cloud Operations Associate & Foundations Associate) |
IBM Developer Skills Network Certifications (Data Science, Machine Learning)