Ashish Kar

Data Science — AI/ML — Data Engineering

+91 8420925607 karashish007@gmail.com linkedin.com/in/AshishKar github.com/AshishKar

Professional Summary

M.Sc. Data Science student with 1 year of hands-on research experience in Machine Learning, Natural Language Processing, and Large Language Models (LLMs). Experienced in building end-to-end AI solutions, creating efficient MLOps pipelines, and deploying models on cloud platforms. I have a strong foundation in explainable AI, deep learning, NLP, GenAI, statistical modeling, time series analysis, and uncertainty quantification.

Experience

Research Intern **Dec 2024 – Present**
BSES Smart Energy Learning Centre *Gandhinagar, Gujarat*

- **Active Learning-Driven NER for Solar News:** Engineered domain-specific NER model on **57K solar news articles with custom 17-entity ontology, achieving 0.83 F1-score improvement using only 8% annotated data** through iterative active learning framework.
- **Sequential Agentic Generation of Knowledge Graphs for Retrieval Augmented Reasoning:** Orchestrated **end-to-end multi-agent agentic Knowledge Graph generation pipeline processing 200+ policy documents and 3 QA datasets**, achieving 90-95% node integration with prompt-optimized entity-relation extraction and cost-reduced orchestration. Developed comprehensive evaluation pipeline benchmarking **agentic KG** construction against **SOTA approaches (OpenIE, GraphRAG, HippoRAG2)** with improved QA accuracy metrics (Paper submitted at **ICDE 2026, Montreal, Canada**).

Technical Skills

Programming Languages: Python, SQL, Basic Bash
Machine Learning & AI: PyTorch, Scikit-learn, Pandas, Optuna, Ollama, LangChain, Hugging Face, CrewAI
MLOps & Big Data: MLflow, DVC, Docker, Git, PySpark, Apache Kafka, Apache Airflow
Cloud Platforms: AWS (S3, EC2), GCP (Cloud Run, BigQuery, GCS)
Coursework: Machine Learning, Deep Learning, Optimization, Statistics, Data Visualization, MLOps, DBMS
Areas of Interest: AI/ML, LLMs, Agentic AI, LLM Personalization, Optimization, Recommendation Systems

Projects

Domain-Adaptive Fine-tuning of Gemma-3 for Financial News QnA | *Unsloth, trl, MLflow, Transformers, LoRA*

- Fine-tuned **Gemma-3 (270M)** on **NewsQA (78k samples)** and **FinancialQA (10k samples)** datasets using **Unsloth** and **trl SFTTrainer**, achieving efficient training through low-rank adaptation and quantization techniques.
- Enhanced model performance while training only about 1% of parameters, raising **F1 from 0.128 to 0.628** and Exact Match from 0 to 0.465 over the pretrained baseline.
- Implemented experiment tracking and hyperparameter management with **MLflow**, and deployed the fine-tuned model on **Hugging Face** for inference, enabling seamless integration with cloud platforms such as **AWS SageMaker** and **Google Vertex AI**.

Credit Card Churn Prediction with Explainable AI | *Python, Scikit-learn, LightGBM, Docker, SHAP, LIME*

- Engineered end-to-end ML pipeline with advanced preprocessing, feature selection, and class imbalance handling achieving **0.89 macro F1-score.**
- Integrated **SHAP and LIME** for comprehensive model explainability, enabling transparent feature impact analysis, prediction interpretation and decision making.
- Deployed scalable real-time inference system on **Google Cloud Run**(earlier on **AWS EC2**) with **FastAPI** backend and **Streamlit UI**, containerized with **Docker** and tested on **70+ users.**

Recommendation System & Analytics Using Beer Reviews | *Airflow, DBT, BigQuery, Glove, MLflow*

- Engineered a scalable ELT pipeline processing **4.5M beer reviews** from PostgreSQL into an analytics-ready star schema in BigQuery, orchestrated with **Airflow** and transformed via **DBT** - ensuring low-latency analytics for large-scale datasets.
- Built a **content-based recommender system** using aspect-based **Glove** embeddings, achieving improved recommendation.
- Integrated **FAISS** for efficient large-scale similarity search, enhancing retrieval speed and quality in the recommender system. Implemented experiment tracking and reproducibility using **MLflow** for model versioning.

Market Basket Analysis & Customer Segmentation Dashboard | *Python, Scikit-learn, Association Mining, RFM Analysis*

- Developed an interactive **customer segmentation and market basket analysis pipeline** using **RFM metrics**, cohort analysis, and association rule mining to uncover key purchasing behaviors.
- Implemented **K-Means clustering** and **RFM-based segmentation** for customer profiling, enabling data-driven retention strategies.
- Applied the **FP-Growth algorithm** from Orange3 to generate high-confidence association rules, enhancing product bundling and cross-selling recommendations.

Eigenportfolio: PCA-based Portfolio Optimization | *Python, NumPy, Pandas, scikit-learn, Matplotlib*

- Developed a PCA-based portfolio optimization framework to construct **eigenportfolios** to enhance diversification and risk-adjusted returns across **419 assets**.
- Constructed diversified, low- concentration portfolios to capture uncorrelated risk factors on historical data from **2000–2013**, using the framework.
- Achieved **72.64% CAGR** with a **1.77 Sharpe Ratio** for the best Sharpe eigenportfolio, demonstrating strong risk-adjusted performance over benchmarks.

BERT-based Review Classifier Using Metric learning | *PyTorch, Transformers, DVC, MLflow, DockerHub*

- Developed BERT-based sentiment classifier with **triplet loss** achieving **0.75 F1-score in 7 iterations** on **4,200** annotated reviews.
- Designed efficient pipeline processing **1.3M Amazon reviews** using statistical sampling under compute constraints. Used **DVC and Mlflow** for dataset and experiment tracking throughout the active learning cycle.
- Implemented dynamic **hybrid sampling strategy in active learning loop** combining **informativeness** and **representativeness** for improved generalization.

Market Structure Analysis of Indian Equities for pair trading | *Scikit-learn, Pandas, Statsmodels*

- Applied **hierarchical clustering on partial correlation matrices** (on the data of **2016-2024**) to uncover systemic co-movements in Indian stock market **(NIFTY50)** and identified statistically significant clusters in IT sector stocks (Monte Carlo p=0.011, modularity ≈ 0.45).
- Validated robustness using bootstrap stability (0.55), temporal out-of-sample testing (ARI=0.22), and cointegration-based filtering of intra-cluster stock pairs.
- Insights applicable to portfolio diversification, sector risk monitoring, and identification of potential statistical arbitrage opportunities.

Granger Causality Between GDP Per Capita - Food Consumption | *Tableau, Python, Statistical Analysis*

- Analyzed **20-year** household consumption data and **10K+ restaurant** records revealing consumption trends and expenditure patterns.
- Established Granger causality relationship between processed food consumption and GDP growth in India using statistical modeling

A/B Testing for Landing Page Optimization | *Python, Statistics, Data Visualization*

- Designed and analyzed an A/B test with 100 users using a **two-sample t-test**, revealing a significant increase in user engagement on the new landing page **(6.22 min vs. 4.53 min, p = 0.00014)**, indicating improved content relevance and usability.
- Validated a rise in conversion rate from **42% to 66%** on the redesigned page through a **two-proportion Z-test (p = 0.008)**, providing data-driven evidence to support adoption of the new design.
- Explored user behavior by language using a **chi-square test**; while not statistically significant (p = 0.21), **EDA** highlighted that French users converted **75% higher** and Spanish users **36% higher**, informing localization and targeting strategies.

Education

M.Sc. Data Science — **2024 – Present**
Dhirubhai Ambani University (Formerly DA-IICT) — *CPI: 9.12/10.0*

B.Sc. Physics Honours — **2020 – 2023**
St. Paul's Cathedral Mission College, University of Calcutta — *CGPA: 8.22/10.0*

Achievements & Recognition

Merit-Based Scholarship: Secured Rank 2 in Semester 1 and Rank 3 in Semester 2, M.Sc. Data Science (2024)
Open Source Contribution: Published 3 Kaggle datasets, 2 Hugging Face models, 4 Docker images, and 6+ GitHub repositories with reproducible code for community use (Dec 2024 - Oct 2025).