ANKIT AHARWAL

+91 7440333991
ankitaharwalrko@gmail.com ◇ Linkedin ◇ Projects

Data Scientist | Generative AI Engineer

PROFILE SUMMARY

IIT Roorkee Alumni with a B.Tech in Computer Science & Engineering and a strong foundation in Data Science.
Over two years of professional experience as a Data Scientist and Software Engineer, with expertise in leveraging data driven insights to drive business impact.
Designed and implemented solutions in data visualization, automation, scraping, large language models and predictive analytics.

SKILLS

Programming & Databases: Python, SQL (MySQL, BigQuery)
Data Science & Analytics: Data Analysis, Statistical Analysis (Scipy, Statsmodels), Feature Engineering, EDA (Autoviz)
Visualization: Plotly, Tableau, Seaborn, Matplotlib, Datadog Dashboards
Machine Learning & Computer Vision: Scikit-learn, XGBoost, PyTorch, Keras, OpenCV, PySpark
LLMs & NLP: LangChain, Chroma DB, GPT-4, OpenAI API, HuggingFace, NLTK, Spacy
MLOps & Data Engineering: AWS Sagemaker, AWS S3, AWS EMR, Airflow, Docker, Kafka, Redis, Elasticsearch
Software Engineering & Tooling: Linux, Docker, Node.js, React, Git, Bitbucket, ETL Pipelines, RESTful API

EDUCATION

B.Tech Computer Science & Engineering, Indian Institute of Technology, Roorkee — May 2018 - June 2022
CGPA: 7.622

Class XII, Narmada Convent School, Barwani — March 2017 - April 2018
Percentage: 88.40%

EXPERIENCE

Data Science Associate (Contract) — April 2025 - August 2025
Ubique Systems — *Pune*

- **End-to-end ML pipeline:** Designed and productionized an XGBoost-based early lung-cancer detection pipeline including feature engineering, k-fold cross-validation, hyperparameter tuning (Bayesian/grid search), and robust class-imbalance strategies (SMOTE & class weights).
- **MLOps & Automation:** Implemented Airflow DAGs for automated ETL, validation, model training, and scheduled retraining triggers, leveraging AWS SageMaker for scalable training jobs and model deployment. Used AWS S3 for model/artifact storage and Bitbucket for CI/CD integration.
- **Model Quality & Monitoring:** Built evaluation and monitoring tooling for AUC/PR-AUC, calibration checks, data-drift detection, and alerting to drive retraining decisions and ensure production reliability.
- **Governance & Reproducibility:** Authored reproducible runbooks, architecture diagrams, and PHI-compliance checklists to meet healthcare governance and audit requirements.
- **Tools:** Python (Pandas, NumPy, scikit-learn, XGBoost), AWS Sagemaker, AWS S3, AWS EMR, MySQL, Bitbucket, Cloudgate.

Data Scientist (Contract) — April 2024 - January 2025
Highbrow Technology — *Remote*

- **RAG & LLM:** Implemented a Retrieval-Augmented Generation pipeline using LangChain Search for semantic retrieval and integrated GPT-4 & Gemini for summarization, Q&A, LLM fine-tuning, and knowledge extraction workflows.
- **Prompt Engineering & Evaluation:** Designed prompt templates, automated prompt-testing harnesses, and evaluation metrics for fidelity, relevance, and hallucination rate reduction.
- **Document Ingestion & OCR:** Built a preprocessing pipeline (OpenCV OCR + text cleaning) to convert scanned documents into indexed embeddings for semantic retrieval.
- **Automation & Tooling:** Developed Python ETL scripts to populate Google Sheets/Drive and created monitoring scripts to validate ingestion quality for downstream RAG.
- **Tools:** LangChain, ChromaDB, GPT-4, Gemini, Python (OpenCV, Pandas), Google Sheets API, Google Cloud.

Software Engineer (Data Engineering Focus) — July 2022 - March 2024
Twilio — *Bangalore*

- **ETL & Automation:** Architected Airflow pipelines to move and transform logs from GCP BigQuery to AWS S3, enabling downstream analytics and ML workflows while reducing manual intervention.
- **Real-time Analytics & Streaming:** Implemented Apache Kafka producers/consumers and PySpark processing jobs for scalable event analytics and near-real-time monitoring.
- **Monitoring & Dashboards:** Designed and maintained 5 Tableau dashboards and Datadog alerts for real-time fraud/activity monitoring, optimized SQL queries for performant visualizations.
- **API Development & Quality:** Designed and implemented RESTful APIs and Golang microservices for data access with ReactJS frontend integration; enforced code quality via Jenkins CI and SonarQube.
- **Performance Optimization:** Tuned Redis usage and implemented cleanup policies to lower cache churn and improve latency.
- **Tools:** SQL, GCP (BigQuery), Python, Tableau, AWS S3, Airflow, RESTful APIs, Golang, Node.js, React, SonarQube, Elasticsearch, Redis, Kafka, PySpark, Datadog, Docker, Jenkins, Git.

Data Science Intern — June 2021 - August 2021
ZS Associates

- **Email Data Project:** Automated email data extraction and applied statistical techniques (A/B testing, CTR, ANOVA).
- **Recommendation System:** Built a pharmaceutical recommendation system with K-Means clustering and Keras.
- **Tools:** Python, Sklearn, Keras, Tensorflow, BeautifulSoup, NLTK, Scrapy, Scipy, and Statsmodels.

PROJECTS

Software Analysis and Design for Trading Cryptocurrency — IIT Roorkee

- Identified profitable cryptocurrency trading strategies.
  Implemented traditional methods (SMAC, EMAC, RSI, MACD, Buy-and-Hold) and ML models (Logistic Regression, Random Forest, SVC, KNN, LSTM, Neural Networks).
- Evaluated performance using a profit-per-day metric with visual comparisons.
  Concluded that SMAC(5,15) and Buy-and-Hold yield the highest median of 4.2 profit per day.
- Technologies: BeautifulSoup (scraping), Fastquant (back testing), Sklearn & PyTorch (ML), Matplotlib (visualization).

Decision Tree Steroid — Data Science Group

- Designed an Advanced Decision Tree capable of utilizing models like SVM, Neural Network, or a standard node.
  Technologies: Sklearn and Matplotlib.

Number Plate Detection — Data Science Group

- Developed a real-time vehicle tracking and number plate detection system.
  Technologies: PyTorch, OpenCV, and YOLO for efficient detection and recognition.

D2l-PyTorch — Data Science Group

- Translated MXNet code into PyTorch for *Dive into Deep Learning* book by Aston Jhang.
  Converted the book into an interactive Jupyter Notebook format and collaborated to maintain d2l-PyTorch on GitHub.
- Technologies: Jupyter Notebook, MXNet, and PyTorch (Deep Learning).

AWARDS

Winner, MIT Covid-19 Challenge

- Team secured 1st position in the MIT Covid-19 Challenge, resulting in $500 prize money.
- Designed an AI-powered IVRS helpline, Suraksha Didi, for offline COVID-19 information access.

POSITION OF RESPONSIBILITY

Data Science Group (DSG), IIT Roorkee

- Led collaborative AI and Machine Learning projects, mentoring peers on data-driven problem solving and research methodology.
- Organized workshops and academic sessions on Data Science and AI, fostering applied learning and interdisciplinary collaboration.
- Contributed to the development of internal curriculum material and open-source resources to support academic learning.