Hrishikesh Keswani

Boston, MA 02135 | (857)3985289 | keswani.hrishikesh7@gmail.com | GitHub | LinkedIn

PROFESSIONAL SUMMARY

Machine Learning Engineer with 2+ years of experience in developing scalable AI platforms and LLM solutions. Expertise in building end-to-end ML pipelines, optimizing inference systems using **PyTorch, LLMs, Ray, Python, AWS.**

EXPERIENCE

Machine Learning Engineer – Boehringer Ingelheim, Athens, Georgia, USA January 2025 – July 2025

- Engineered scalable **RAG** architecture using **Llama-3**, **GPT-4o**, and **Nova Lite** that automatically extracted key themes from 10k+ employee survey responses, enabling data-driven leadership decisions through contextually-aware response generation
- Improved chatbot throughput by **80%** through asynchronous **FastAPI** endpoints and **vLLM** continuous batching, containerized via **Docker** and scaled across multi-GPU clusters using **Ray** for parallel workload distribution
- Developed an automated **AirFlow** data engineering pipeline (scraping, cleaning, chunking, and vector embedding) with nightly updates and drift-triggered retraining, ensuring high-quality model inputs and reducing inaccuracies
- Leveraged **AWS CloudWatch**, **Grafana**, and **Prometheus** to design comprehensive monitoring and alerting systems, decreasing incident resolution time by **40%** and achieving **95%** SLA compliance

Data Scientist - Yunometa, Mumbai, India August 2022 - July 2023

- Analyzed sales and product pricing data from SAP of 2 million + records to provide actionable insights by creating various visualization dashboards in **Power BI** and **SAP Analytics Cloud**
- Designed an automated **ETL** (Extract, Transform and Load) pipeline using Python and MSSQL that processed and analyzed machine data, resulting in a **30%** reduction in manual data processing time
- Developed customized, enterprise-level data visualization dashboards in **Tableau** for Amazon Product Sales and Ecommerce Profit and Loss, utilized by various departments including Amazon Operations, Revenue Management, Marketing and Logistics
- Analyzed complex data and examined industry standards to help the FP&A team to determine and set competitive prices

Data Scientist - Kritexco, Mumbai, India December 2021 - March 2022

- Analyzed 50k+ dataset from OpenSea and Rarible using Python and **Matplotlib**, extracting insights on NFT sales trends. Visualized findings in **Tableau**, resulting in better decision making
- Built and deployed an **LSTM** model using **Tensorflow** for predicting NFT prices with an RMSE score of **386**, resulting in a **20%** improvement in price accuracy compared to traditional methods
- Collaborated closely with stakeholders and cross-functional teams to identify data requirements, design scalable data pipelines, and ensure data accuracy and integrity through development, test, and production environments

Machine Learning Engineer- Aadyam Infotech, Pune, India June 2021 - August 2021

- Formulated **regular expressions**, leveraged **Azure AI**, and engineered an NLP pipeline that parses and extracts key information from unstructured documents, and achieved an accuracy of **86%** for extracted records
- Worked with cross-functional teams in an agile environment to develop and brainstorm optimization strategies for the algorithm

PROJECT EXPERIENCE

FinanceHelper AI | GenAI | NLP | Retrieval Augmented Generation | PEFT | LLM September 2024

- Fine-tuned a Large Language Model (**Mistral-7B**) language model on the book "The Psychology of Money" with **LoRA**, and **NLP** techniques for tokenization and integrated vector dbs, further improving the accuracy and speed of the model's responses
- Integrated Retrieval-Augmented Generation (**RAG**) using **LangChain** and semantic similarity, fetching relevant advice from the book pertinent to the user's query, improving the accuracy and contextual relevance of the response

LLM-Driven Financial Sentiment Analysis Tool | LLM | Data Mining | NLP | AI | Prompt Engineering May 2024

- Utilized Python's **Scrapy** for web scraping of financial news, integrating **BERT** transformers to perform sentiment analysis and text summarization and achieved a **90%** accuracy, boosting strategy performance
- Fine-tuned the **BERT** model on a custom dataset of financial news articles, using learning rate scheduling, gradient clipping, and data augmentation to improve sentiment analysis accuracy

Prediction Analysis of Diabetes Patients Readmission | Machine Learning | Decision Trees | Predictive Modelling March 2024

- Developed **LSTM**, **CatBoost**, and **Random Forest** models to predict readmission rates from a **100K** record EMR dataset
- Applied feature engineering and statistical tests (**Chi-square**, **Spearman**) to optimize data preprocessing, reducing training time by **30%**
- Achieved F1 scores of 0.68, 0.72, and 0.89, with **CatBoost** performing best (Precision: 0.72)

EDUCATION

Northeastern University | Master of Science in Data Analytics Engineering December 2025

Relevant Coursework: Foundations for Data Analytics, Data Management for Analytics, Data Mining, Computation & Visualization, Neural Networks & Deep Learning, Natural Language Processing, Database Management Systems

TECHNICAL SKILLS

Core ML/AI	PyTorch, LangChain, HuggingFace, LLMs (Llama-3, GPT-4o, Mistral), vLLM, MLFlow
Programming	Python, SQL, Java, C++, Golang, TypeScript, Scala
Databases	Apache Spark, Airflow, Kafka, Redis, PostgreSQL, MongoDB, Druid, Solr
Cloud & DevOps	AWS (S3, Lambda, EC2, EKS, EMR), Docker, Kubernetes, OpenShift, CI/CD
APIs & Frameworks	FastAPI, Django, GraphQL, TensorFlow/Keras, Springboot