RISHIKESH DHAYARKAR

San Jose, CA | github.com | linkedin.com | medium.com | rishikeshdhayarkar1091@gmail.com | 347-471-8506

Data Scientist with **4 years** of expertise in delivering production-grade ML and analytics solutions. Skilled in leveraging ML paradigms, LLMs, data & software engineering practices to develop products that solve complex data and AI challenges.

EDUCATION

NEW YORK UNIVERSITY *New York City, NY*
Master of Science in Computer Science and Engineering *Aug 2019 - May 2021*

VISVESVARAYA TECHNOLOGICAL UNIVERSITY *Bangalore, KA*
Bachelor of Engineering in Electronics and Communication *Aug 2014 - May 2018*

WORK EXPERIENCE

DATA SCIENTIST *(ML and NLP)* **at FRACTAL** *Sep 2021 – present | San Jose, CA*
- Developed an LLM-powered NLP application for rapid text analytics. Enabled UX researchers and analysts to extract insights from text data stored in csv files and SQL tables in minutes. Leveraged Python, Gemini models/APIs, TensorFlow, Flask, FastAPI, Hugging Face, and GCP to build the application. Implemented features including topic modeling, topic discovery, sentiment detection, text summarization, and question answering.
- Built LLM classifiers to identify product issues and user feedback with an average F1 score of 96%. Built classifiers by fine-tuning Gemini models with LoRA, few-shot learning, and advanced prompt engineering. Utilized Python, Gemini APIs, Hugging Face libraries, and active learning for dataset construction.
- Developed classification models for a leading video sharing platform, resulting in a 14% increase in Revenue over Ad Spend (ROAS) for the marketing team. Achieved this by constructing a customer sign-up propensity model (ROC AUC = 0.98, precision-recall AUC = 0.97) and a subsequent spend tier prediction model. Leveraged Python, TensorFlow, TFX, and BigQuery for model building, deploying, and setting up ETL pipelines for distributed data ingestion, cleaning, and preprocessing.
- Collaborated with UX teams to identify inefficiencies in the sign-up funnel of a video-sharing platform. Conducted A/B tests on funnel improvements, resulting in an 8% increase in customer acquisitions and an 11% reduction in sign-up time.
- Constructed ETL pipelines leveraging SQL, BigQuery and crafted a dashboard using Looker to monitor Ad campaign health. Created timely campaign performance reports/presentations with statistical and business insights.

DATA SCIENTIST *(NLP)* **at PROMAZO** *Aug 2021 – Oct 2021 | Chicago, IL*
- Designed a similarity search-based model for a start-up, resulting in a 40% improvement in matching accuracy for their consumer product connecting students with industry experts for personalized career mentorship. Leveraged Python, Pytorch, Pandas, Lambda, EC2, S3, and HuggingFace, for feature extraction, transformer based embedding generation and model building and deployment.

RECENT PROJECTS (ALL PROJECTS)

PODPROBE: PODCAST EXPLORATION TOOL | Github | *OpenAI API, Pinecone, Neo4j, LlamaIndex* *Jul 2024 – Present*
- Built a dual Retrieval-Augmented Generation (RAG) system, combining traditional and graph-RAG for fast fact retrieval and complex entity mapping, enabling efficient podcast information extraction.
- Integrated the system into a ReAct agent, handling multi-hop queries efficiently, reducing RAG calls, and minimizing latency. Utilized OpenAI API for audio transcription, Neo4j for knowledge graphs, Pinecone for vector storage, and LlamaIndex for agent development.

ML-OPS PIPELINE FOR HATE SPEECH DETECTION | Github | *Pytorch, Docker, FastAPI, Google Cloud Platform* *Apr 2024 – Jul 2024*
- Built a production-grade MLOps pipeline for detecting hate speech on online platforms. Leveraged Docker for reproducibility, Dask for distributed data processing, and PyTorch Lightning for distributed model training, with deployment on Google Cloud Platform (GCP).
- Developed a scalable, maintainable, and cloud-ready infrastructure, serving the model via a REST API using FastAPI and Streamlit.

ML MODEL API FOR HOUSE PRICE PREDICTION | Github | *Python, FastAPI, Docker, Railway, Scikit-learn* *Feb 2024 – Apr 2024*
- Developed a house price prediction API using FastAPI, on a regression model trained on Kaggle's advanced house pricing dataset.
- Implemented end-to-end ML ops pipeline, including containerization with Docker and deployment on both PaaS (Railway) and IaaS (AWS EC2) platforms. Set up CI/CD with CircleCI, automated testing across environments using Tox, and package distribution via Gemfury.

NYC TAXI DATASET ANALYSIS | Hadoop - Github | Spark -Github | *python, pyspark, hadoop* *Nov 2020 – Dec 2020*
- Performed comprehensive EDA and identified data quality issues in the NYC Taxi Dataset.
- Used Spark and Hadoop on the NYU HPC cluster to facilitate efficient computation. Dataset size - 4GB

SKILLS

PROGRAMMING : Python, Java, Javascript, SQL, HTML, CSS, R
ML-OPS AND DEV TOOLS : Docker, Git, Weights & Biases, MLFlow, Circle CI, Github actions, TorchServe, Flask, FastAPI, Google Cloud Platform, AWS, Spark MLLib, PySpark, SparkSQL, BigQuery, Neo4j, Pinecone, DVC, Hydra-core
ML FRAMEWORKS AND TOOLS : Pytorch, TensorFlow, HuggingFace, OpenAI API, Gemini API, LlamaIndex, Langchain, Scikit-learn, Pandas, OpenCV, Nltk, Numpy, Gensim, XGBoost, LightGBM
ANALYTICS : Hypothesis Testing, Time-series Analysis, Causal Impact Analysis, A/B Testing, Statistical Inference, Excel, Statsmodels