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. Proficient in building ML models leveraging best Data/Software Engineering practices and Statistical Analysis to tackle complex business 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 ***(Behavioral Analytics | NLP)*** **at FRACTAL** — ***Jun 2023 – present | San Jose, CA***

- Data Scientist on the Behavioral Analytics team, analyzing user perceptions to guide ad strategy and product development.
- Created end-to-end text analytics workflows to detect product issues and feature requests from reviews, surveys, and feedback from millions of users. Organized product issues into taxonomies, recommended new product features and design optimizations for UI.
  * Developed LLM classifiers (avg. F1 score: 0.96) by fine-tuning Gemini models using Low Rank Adaptation(LoRA) and advanced prompt engineering. Leveraged Python, Gemini APIs, Hugging Face libraries, SQL and active learning tools for development.
  * Transitioning to LLM classifiers improved F1-scores (+30pp), eliminated manual checks & boosted stakeholder trust.
- Developed an LLM-powered application for efficient text analytics, allowing internal teams of UX researchers/analysts to extract insights from CSVs and SQL tables in minutes.(250+ active internal users/month)
  * Implemented key features like topic modeling, topic discovery, sentiment/emotion detection, and summarization leveraging Python, Gemini models/APIs, TensorFlow, Flask, FastAPI, Hugging Face, and GCP.
  * Ensured code quality with Test-Driven Development (TDD), CI/CD, and created low-code notebooks for easy adoption by analysts.

DATA SCIENTIST ***(Product Marketing Analytics)*** **at FRACTAL** — ***Sep 2021 – Jun 2023 | San Jose, CA***

- Product Marketing Data Scientist at a leading video-sharing platform, guiding the paid media team in optimizing marketing strategies.
- Developed neural network based classification models resulting in a 14% increase in Revenue over Ad Spend (ROAS). 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 (ROC AUC = 0.90, precision-recall AUC = 0.89). Leveraged Python, tensorflow, TFX for development.
- Designed and implemented SQL-based ETL pipelines to process data from 60 million customers, handling tens of TBs of data. These pipelines orchestrated distributed ETL processes, implementing data quality checks, schema evolution, and partitioning strategies to optimize large-scale data preparation for ML model ingestion.
- Collaborated with UX teams to identify inefficiencies in the customer sign-up funnel. 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 and crafted dashboards using Looker to monitor Ad campaign health KPIs. Analyzed the effectiveness of advertising campaigns by implementing hypothesis testing, causal inference, and time series analysis.

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)

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 classifying 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.

SKILLS

PROGRAMMING : Python, SQL, Java, Javascript, HTML, CSS, R
ML-OPS : Docker, Git, Weights & Biases, MLFlow, Circle CI, Github actions, Flask, FastAPI, GCP, AWS, PySpark, SparkSQL, BigQuery, DVC
ML FRAMEWORKS : Pytorch, TensorFlow, HuggingFace, LlamaIndex, Scikit-learn, Pandas, Nltk, Numpy, Gensim, XGBoost, LightGBM
MODELING : K-Means Clustering, Hierarchical Clustering, DBSCAN, Logistic Regression, Decision Trees, Random Forest, Support Vector Machines (SVM), K-Nearest Neighbors, Neural nets, Linear Regression, Decision Tree Regression, Random Forest Regression, ARIMA, Prophet
ANALYTICS : Hypothesis Testing, Time-series Analysis, Causal Impact Analysis, A/B Testing, Statistical Inference, Excel, Statsmodels