RISHABH INDORIA

732.486.6359 | indoria.r@northeastern.edu | LinkedIn | GitHub | Portfolio

EDUCATION

Northeastern University Boston, MA
Master of Science in Information Systems (Teaching Assistant – Advances in Data science and Architecture) Expected April 2024
Manipal Institute of Technology (MIT) Manipal, Karnataka, India
Bachelor of Technology in Information Technology 2013 - 2017

SKILLS

- **Software Programming and Shell Scripting:** Python, Spark, SQL, Kafka, Algorithms, Linux, Data structures, Algorithms, APIs
- **Machine Learning:** Supervised learning, Scikit-Learn, NLTK, HuggingFace, Statistical modeling, OpenAI, PyCaret
- **Deep Learning:** TensorFlow, Keras, PyTorch, Computer vision, Unsupervised modeling, RNN, CNN, GAN, GNN, Transformers
- **Cloud computing:** AWS (Ec2, IoT, S3, Lambda, Redshift), GCP, Git, Airflow, Terraform, SageMaker, Docker
- **Data Warehouse:** Snowflake, Redis, Pinecone, InfluxDB, Telegraf, Hadoop, Databricks, Postgres, MongoDB
- **Data Visualization and Business Intelligence:** PowerBI, Tableau, Grafana, Excel, SAP Analytics cloud

WORK EXPERIENCE

URJA.IO | ***Lead Data Scientist*** April 2020 - August 2022

- Spearheaded development team in creating an end-to-end analytical dashboard tool, resulting in a 25% increase in user retention
- Reduced costs by 23% with a 94% accurate machinery-failure **prediction system** using **Prophet** for time-series **forecasting**
- Streamlined stakeholder communication by 10 hours weekly through efficient **Tableau reporting** of **A/B test** progress
- Collaborated with **cross-functional** teams to implement pricing strategies, contributing to overall business roadmap planning

SOCIETE GENERALE GLOBAL SOLUTION CENTRE | ***Data Scientist*** July 2017 - March 2020

- Enhanced Infrastructure as code (IaC) ETL pipeline deployment speed by 30% through Terraform and GitHub **CI/CD** optimization
- Increased **anomaly detection** in banking by 17% using **LightGBM** models for identifying irregular financial transaction volumes
- Boosted **risk management** efficiency by 21% using **Isolation Forest** to spot high-value transactions exceeding forecasted pricing
- Employed **NER** and **POS** tagging techniques to redact financial transcripts, achieving a low error rate of 3.5%

PROJECTS

IMBALANCED METEOROLOGICAL DATA ANALYSIS FOR RAINFALL PREDICTION | **GITHUB** December 2023

- Enhanced prediction accuracy in an **imbalanced dataset** using **SMOTE, class weight adjustments, and optimizer tuning**
- Developed and trained multimodal **ConvLSTM** model for Rainfall Forecasting Achieving **77% Class-1 Accuracy**
- Trained integrated **RNN-TimeDistributed CNN** model for dual time-series and image data analysis, achieving **93% F-1 score**

IMAGE SEGMENTATION WITH CONTEXTUAL CAPTIONING | **GITHUB** October 2023

- Implemented **CNN** with **PyTorch** and **OpenCV** for efficient image segmentation and **region of interest (ROI)** with 0.92 F1-Score
- Developed **RNN** in **TensorFlow** and **NLTK** for generating context-aware captions from segmented data with 70 **Perplexity** score

ENHANCED AUDIO JOURNALING WITH LLM | **GITHUB** August 2023

- Implemented containerized **audio processing** algorithms, focusing on emotion detection for accurate **sentiment analysis**
- Leveraged **Snowflake** for efficient and scalable data storage, facilitating data integrity and robust data management
- Utilized GPT-3.5 **prompt engineering** for generating personalized feedback, achieving **20% improvement** in user insight

FINANCE SEARCH OPTIMIZATION | **GITHUB** May 2023

- Engineered a Q&A and **text summarization** platform, ingesting streaming transcripts from **Kafka pub-sub** and storing it in **Redis**
- Used **Langchain** traditional filters, RAG and **Pinecone** for vector similarity to improve information retrieval quality by 50%
- **Dockerized Airflow** data pipelines and **Spark** processing strategies, resulting in a **50% reduction in information retrieval time**

SENTIMENT ANALYSIS: DECIPHERING RESTAURANT REVIEWS | **GITHUB** April 2023

- Developed a dataset of restaurant reviews through **prompt engineering** using **OpenAI's** LLM Davinci model
- Applied text preprocessing techniques (**tokenization, lemmatization**) followed by **TF-IDF Vectorization** and **Naive Bayes** algorithm for sentiment analysis on real-time data, achieving a classification accuracy of 91%

ECONOMICS OF HAPPINESS | **GITHUB** November 2022

- Utilized **logistic regression** and **decision trees** for exploratory analysis of the relationship between economic indicators and happiness index across 150 countries, enhancing model insights by 15% through **hyperparameter tuning** and **SHAP analysis**

ACHIEVEMENTS

- Presented at AI Skunkworks' "NLP Deep Dive" technical workshop with the Institute for EAI (Watch Video)
- **Microsoft** Certified: Azure Data Scientist Associate: Certificate