Shubham Shailesh Tamhane

shubhamtamhane2000@gmail.com linkedin.com/in/shubhamtamhane
github.com/shubhamtamhane shubhamtamhane.github.io

Education

University of Rochester **Aug 2022 – Dec 2023**
Master of Science in Data Science *Rochester, NY*

- GPA: 3.96/4. Recipient of 40% merit scholarship
- Secured 2nd position in the 2022 UR Biomedical Data Science Hackathon

University of Mumbai **Aug 2018 – May 2022**
Bachelor of Engineering in Information Technology *Mumbai, India*

- GPA: 3.73/4, CGPA: 8.95/10
- Hackathon winner and Guest speaker for creating video conferencing web application (Google Meet clone)

Relevant Courses: Time Series, Data mining, Statistics, NLP, Machine Learning, AI, Data Structures, Big Data, DBMS

Technical Skills

Programming Languages: Python, SQL, R, C, C++
Databases: Microsoft SQL Server, MySQL, PostgreSQL, MongoDB, NoSQL, SSMS, S3, Redshift, DynamoDB
ML Domains: Regression, Classification, Clustering, Natural Language Processing (NLP), Generative AI (GenAI)
Libraries and Services: Pandas, Numpy, Matplotlib, Scikit-Learn, PyTorch, TensorFlow, AWS (Certification)
Analytics Tools: PowerBI, Tableau, Excel, Microsoft Office, Apache Airflow, Git, Docker, MLFlow

Experience

Indiana University **Mar 2024 – Present**
Data Scientist *Bloomington, IN*

- Incorporated **Microsoft SQL Server** in combination with **Python** for in-depth analysis of healthcare datasets, exceeding **5M** records, to derive insights from Electronic Health Record (**EHR**) data.
- Enhanced data quality by applying feature engineering, outlier detection, and missing value imputation using advanced data preprocessing techniques in Python (Pandas, NumPy).
- Developed a classification model for disease prediction with recall of 85% by implementing **XGBoost** in **scikit-learn**.
- Collaborated with researchers to predict survival activity, improving prediction reliability by 30% by implementing deep learning architectures in **PyTorch** trained on patient's food intake and activity records.

Regeneron Pharmaceuticals **Jun 2023 – Dec 2023**
Data Science Intern *Tarrytown, NY*

- Implemented time series forecasting approach to predict customer demand of a complex **inventory management** problem employing multiple approaches including **statistical** and **deep learning** methods.
- Deployed a webapp built using **python-dash** that leverages **MLOps** workflow built on cloud-infrastructure to provide real-time up-to date data and forecasting predictions, customer analysis and model maintenance options to end users contributing significantly to **cost optimization**.
- Led the development of a **maintenance analysis** system, optimizing the upkeep of MFCs and related systems, which resulted in substantial monthly savings.
- Adopted **JIRA** for task tracking and **Confluence** for documentation, adhering to the **Agile/Scrum** methodology.

Zalliant **Sep 2023 – Dec 2023**
Data Scientist *Amsterdam, NY*

- Led a team of **3** data scientists to extract features representing the current activity of cows from video data using **Excel** and **Boris** software, catering to multiple targets within the same frame.
- Implemented **Random Forest** as a MultiOutputClassifier, achieving a 97% accuracy and 92.43% F1 score by extracting time and frequency domain features for behavior classification.
- Engineered optimization strategies resulting in a **99.6%** reduction in sensor data collection frequency, thereby extending sensor battery life by **3** months.

URMC - Center for Advanced Brain Imaging and Neurophysiology **Sept 2022 – Jun 2023**
Software Intern *Rochester, NY*

- Created a distributed **ETL** pipeline to process over 10,000 DICOM files, reducing ingestion time by 50% by leveraging **Python (PySpark)** for metadata extraction and storing structured data in **MySQL**.
- Orchestrated automated data validation workflows, enabling ingestion of new DICOM files by employing **Apache Airflow** to enforce data integrity checks before storage.
- Containerized the data pipeline, enabling cross-platform execution and reducing deployment time by 80% using **Docker**

Sciffer Analytics Pvt Ltd **Oct 2020 – Jan 2021**
Data Science Intern *Pune, India*

- Managed the development of image datasets using labelimg tool for **information extraction** from Google in 3 months empowering a computer vision model to recognize over 30 distinct objects.
- Employed the **YOLO v3** model to build a deep learning classifier model, attaining an accuracy rate of **80%**.

Projects

RAG-Based QA with Langchain and OpenAI | Project Link **Jun 2024**

- Engineered a context-specific question-answering system leveraging the Python REPL tool and **FAISS** vector database for embedding and retrieval operations with max-marginal-relevance strategy.
- Utitlized **Streamlit** for creating a user interface, facilitating file uploads and invoking RAG chains for interactive Q&A sessions using **OpenAI GPT Instruct** models.

Document Query System Using LLama2 and LlamaIndex | Project Link **Dec 2023**

- Created a question-answering system for domain-specific PDF documents by implementing **RAG** with quantization of open source **LLama2** model for enhancing computational efficiency.
- Utilized **LlamaIndex** framework and **VectorStoreIndex** to store embeddings in combination with **HuggingFace**, enabling accurate and efficient domain-specific query responses.

Dynamic QA generator for Research Papers | Project Link **May 2023**

- Fine-tuned a T5-base model and integrated **GenAI** to create a **Question-Answer** system that generates and answers questions from research papers, enhancing paper interpretation.
- Employed the **QASPER** dataset to evaluate models, achieveing a **BLEU** score of 0.85, **ROUGE** score of 0.78, and **QAeval** score of 90%.

Emotion Recognition Using Deep Convolutional Neural Networks | Publication Link **Apr 2022**

- Neural networks such as **ResNet50** and **VGG16** were used to identify the mood of the user based on facial expression.
- Applied **Haar Cascades** on the FER2013 dataset, followed by a custom deep convolutional neural network (DCNN) to achieve an accuracy of **83.9%** .