Shubham Shailesh Tamhane

+1 (585)-363-9693 | shubhamtamhane2000@gmail.com | LinkedIn | Portfolio | GitHub

WORK EXPERIENCE

Data Science Analyst *03/2024 – Present*
Indiana University *Remote, USA*

- Applied statistical methodologies to extract insights from Electronic Health Record (**EHR**) data using **T-SQL** and **Python**
- Developed **ETL** pipelines to automate the ingestion, transformation and loading of Microsoft SQL Server data exceeding **50M** records, thus reducing processing time by **30%**
- Implemented data validation and quality assurance protocols across multiple sources and systems of healthcare data while adhering to **HIPAA** compliance standards
- Built a classification model using **XGBoost** to predict disease outcomes in a heavily imbalanced dataset, integrating EHR and socio-economic external variables, achieving an **AUC** score of **0.76**
- Created predictive models (LSTM, RNN) to assess disease diagnosis over 10 years of using **Tensorflow** and **Pytorch**
- Wrote **20+ SQL** scripts to transform raw data into structured formats suitable for dashboard integration
- Designed interactive dashboards in **PowerBI** to visualize the demographic spread of disease across USA

Data Scientist Intern *06/2023 – 12/2023*
Regeneron Pharmaceuticals *Tarrytown, USA*

- Developed ETL pipelines in **AWS Glue** for time series forecasting on a quarterly interval for an inventory management system stored in **Amazon Redshift** with **30K+** products
- Built time series models (ARIMA, ETS) in AWS **SageMaker** to forecast product demand, achieving **80%** accuracy
- Reduced material waste by **30%** per quarter through improved supply chain efficiency and forecasting
- Engineered an end-to-end **MLOps** pipeline using **Dash** and **mlflow** to enable real-time predictions, customer analysis and model retraining workflows
- Utilized Amazon **Quicksight** to design interactive dashboards and alarms by fetching data from **OSI PI**, thus optimizing workflows and reducing task turnaround time for end users by **20%**
- Adopted **JIRA** for task tracking and backlogs and **Confluence** for documentation, adhering to the **Agile/Scrum** methodology

Data Scientist Consultant *09/2023 – 12/2023*
Zalliant *Remote, USA*

- Developed a scalable machine learning pipeline on **Databricks** using medallion architecture and **Spark**
- Built a multiclass classification model to distinguish cattle behavior achieving **97%** accuracy and **92.43%** F1 score
- Utilized Random Forest as a MultiOutputClassifier, extracting Time and Frequency domain features for improved cattle management decision-making
- Performed exploratory data analysis (**EDA**) and data cleaning using **pandas**, **numpy** and **matplotlib** to ensure data quality for machine learning model development, thus reducing model training time
- Conducted **A/B testing** and statistical analysis to validate model assumptions, resulting in increase in the confidence level of the cattle behavior classification model's real-world effectiveness

Software Engineer *09/2022 – 05/2023*
URMC – Center for Advanced Brain Imaging and Neurophysiology *Rochester, USA*

- Instituted a verification system using **pydicom** to confirm over **10,000** DICOM files within 60 seconds
- Utilized MongoDB to store metadata of DICOM files for data management and retrieval
- Built a **Flask** web service with JWT authentication and caching, reducing large JSON file load times to under 10, thereby improving load time by **91.6%**
- Deployed the web application using **Docker** containers for isolated execution across 3 operating systems

Data Scientist Intern *10/2020 – 01/2021*
Sciffer Analytics *Remote, India*

- Created and annotated image datasets using **labelimg**, extracting images from Google to build training dataset
- Implemented **YOLOv3** model and custom CNNs to detect objects in image pauses and frequently viewed video segments
- Performed analysis of video engagement metrics, leveraging object detection insights to understand viewer behavior

SKILLS

Programming Languages: Python, SQL, R, C++, Javascript

Databases: PostgreSQL, DynamoDB, MongoDB, MS SQL Server, MySQL

Frameworks: PyTorch, TensorFlow, Keras, Sklearn, Django, Flask, Node.js

Data Management and Analytics tools: Apache Airflow, Spark, PowerBI, Tableau, Excel

CI/CD and DevOps: Docker, JIRA, Git, Bitbucket, Kubernetes

Cloud Platforms: **AWS Certified Cloud Practitioner**, Azure, Databricks, Snowflake, Dataiku, Seeq

EDUCATION

University of Rochester *08/2022 – 12/2023*
Master of Science in Data Science (GPA: 3.96/4) *Rochester, USA*

- **Coursework**: Statistics, Data Mining, Time Series, NLP, Machine Learning, Big Data, DBMS, Business Intelligence